Catalogue No. 93-03A

SLID MICRODATA FILES
CONTENT PROPOSAL
PART A - OVERVIEW

June 1993

Jamie Brunet, Household Surveys Division

Philip Giles, Household Surveys Division

The SLID Research Paper Series is intended to document detailed studies and important decisions for the Survey of Labour and Income Dynamics. These research papers are available in English and French, at no charge. To obtain a summary description of available documents or to obtain a copy of any, please contact Philip Giles, Manager, SLID Research Paper Series, by mail at 11-D8 Jean Talon Building, Statistics Canada, Ottawa, Ontario, CANADA K1A 0T6, by telephone (613) 951-2891, or by fax (613) 951-3253.

EXECUTIVE SUMMARY

In February 1992, the SLID (Survey of Labour and Income Dynamics) Content Development team widely distributed a proposal for survey content. Using this document as the basis for discussion, extensive consultation resulted in revisions to the survey content.

A similar approach to SLID output products is planned. This report (with Parts A and B bound separately) is intended as an initial proposal for SLID longitudinal microdata files. Feedback is strongly encouraged. Any suggestions for other SLID output products are welcome. Your comments can be sent to the Manager, SLID Research Paper Series -- contact information is provided on the cover page.

This document, Part A, provides an overview of the strategy proposed for SLID longitudinal microdata files. Part B, available upon request, will be of interest to those wanting detailed information on specific data variables.

While suggestions and comments will be welcome at any time, a deadline of October 31, 1993 is requested. At that time, a revised strategy will be developed, and made available through the SLID Research Paper Series.

TABLE OF CONTENTS

1. OVERVIEW OF SLID DISSEMINATION

To date, the SLID (Survey of Labour and Income Dynamics) team has focused on survey content and data collection. While these aspects are important, many users are naturally very interested in our output plans. To meet the needs of data users, SLID plans a range of products, and seeks advice in developing them. This document is the first step in this endeavour. The first SLID data will be released in early 1995. Throughout 1994, output plans will be developed and refined.

Until data become available, the SLID product line includes two items. A newsletter, *Dynamics*, provides updates on survey developments and issues as they arise. Distributed four times annually, the newsletter keeps readers broadly informed about the survey.

A SLID Research Paper Series exists for those interested in following developments more closely, including survey design issues, data quality evaluation and exploratory research. Both the newsletter and research papers are available at no charge. A person may subscribe to all research papers, or order individual copies of those of interest only. Every research paper is briefly described in *Dynamics*.

Overview of Product Line

The main benefits of a longitudinal survey are derived from the analysis of microdata, specifically of changes over time of particular characteristics at the individual level. Therefore, the provision of microdata and good documentation is our first concern, and is the primary focus of this document.

Various microdata products (discussed in more detail in Section 4) will be offered. If there is sufficient demand, SLID data user workshops will be held to help with the understanding and use of the data.

The product line will also include analytical publications, concentrating on topics exploiting the use of SLID data. Our early thinking on the SLID team's role with respect to analysis may be characterized as follows:

- SLID will perform analyses as part of the data release process. Although data release should be as timely as possible, it should be accompanied by a substantial piece of analysis. To do less would underrate the data.

- Depending on funds, research contracts will be used to promote studies illustrating data applications and analytical techniques particularly suited to SLID. Results will be published by Statistics Canada, with full credit to the authors.

- We will look for opportunities to participate in joint analysis projects with researchers, both inside and outside Statistics Canada, as this increasingly appears to be an excellent route to getting the best of both worlds.

- SLID will participate in, and perhaps fund, analysis projects intended to use SLID data in conjunction with other longitudinal survey data.

Although there are no firm plans to do so, standard data publications will be produced if a common set of tabulations with wide appeal can be identified. The difficulty is how to highlight longitudinal aspects of the data. Indeed, the survey's content is similar to current cross-sectional surveys (like the Labour Force Survey and the Survey of Consumer Finances), which have well-developed data

publication programs. SLID's objective is to complement these existing surveys, and it's not clear how data publications would achieve this goal. Moreover, time spent by staff on the production of data publications would reduce time spent on the production of microdata products and associated documentation.

2. PURPOSE AND SCOPE OF DOCUMENT

A document describing the proposed SLID content was widely distributed in February 1992. The intention was to obtain feedback from potential data users before proceeding with questionnaire development. This document was prepared in the same spirit: to solicit advice on how best to structure microdata products before developing production systems. Part A provides an overall description of the proposed SLID longitudinal microdata files; Part B provides detailed specifications for every variable.

This report is designed to be a starting point. This means, first, that very little is firmly decided. Second, it does not cover the full product line; future documents will incorporate changes resulting from this consultation and will raise new issues. The enormity of the task of defining output products seemed to lead to this "progressive" type of consultation, rather than waiting for all aspects to be fully investigated prior to involving data users.

This report includes:

- an explanation of the general structure of data variables;
- the proposed frequency of release;
- a description of derived variables (in Part B).

This report does not include:

- detailed algorithms for derived variables, which will be produced at a later stage;

- final survey content -- the report lists content as currently known;

- a complete airing of the confidentiality issue, although a brief discussion is included;

- the method to be used for indicating changes made to the data during processing. SLID microdata products will identify imputed values, but no method of doing so has yet been proposed.

3. CONFIDENTIALITY

In Part B of the report, all variables are presented in their unscreened forms. It is clear that some variable suppression and collapsing of values will be necessary before releasing any public-use microdata products. On the other hand, Statistics Canada recognizes that certain analyses will not be feasible using screened microdata. A mechanism will be developed to allow researchers to use the complete range of microdata for inputs into statistical analyses, without jeopardizing the confidentiality assurances provided to respondents. (A future SLID Research Paper will discuss this issue.)

Although details are far from final, the following scenario is envisioned:

- Public-use microdata products will contain all data variables. Where screening is necessary to protect confidentiality, a technique other than suppression will be used -- for example, assigning feasible values to

individuals such that mean values are maintained within certain population classes.

- If researchers would find it useful for some initial exploratory work, it might also be possible to include complete data (i.e., no suppression or collapsing of values) for a few hundred individuals to use as an indication of the possible range of values available. Obviously, the individuals would have to be carefully chosen to prevent the actual identification of the person from the data.

- For studies requiring unscreened data, the researcher would write the code to do the data extraction and analysis, and telecommunicate the code to Statistics Canada. The program would be run using the in-house data file. After verifying that nothing in the output contravenes confidentiality considerations, the results would be transmitted to the researcher. To the extent possible, the entire process would be automated; the greater the automation, the faster the turnaround time and the lower the cost to the researcher.

4. FREQUENCY OF FILE RELEASES

SLID intends to release two types of files: cross-sectional and longitudinal. A cross-sectional file contains information for one reference year only, and will be similar to a "traditional" survey microdata file. Due to its unique content mix of labour and income, the SLID cross-sectional file will be a major product in itself, as it will provide data not currently available from other sources. Longitudinal files, the primary focus of this document, cover more than one reference year.

Table 1 illustrates the relationship between the reference periods for data collection and the proposed file releases. One cross-sectional file will be released annually, providing all SLID information for a particular reference year. Due to the staggered introduction of the sample, cross-sectional files covering the first three years of data collection will contain data collected from one panel of respondents. Starting with reference year 1996, cross-sectional data will be based on two panels of respondents. (The initial sample for each panel is approximately 20,000 households.)

TABLE 1 - Microdata File Availability, by Type

	Reference Year											
	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
File X	●	●	●	●	●	●	●	●	●	●	●	●
File L3	<=	==	=>	<=	==	=>	<=	==	=>	<=	==	=>
File L6	<=	=*	==	=*	=*	=>						
File L6				<=	==	==	==	==	=>			
File L6							<=	==	==	==	==	=>

Note: File X = cross-sectional file; File L3 = 3-year longitudinal file;
File L6 = six-year longitudinal file; * = exceptions to proposal

Due to the nature of the sample design (overlapping panels) and panel rotation, longitudinal files covering two time periods are proposed: three-year files and six-year files. The three-year files will be based on the sample for two panels (exception: the first will be based on one panel due to the phased introduction of

the sample). This larger sample will suit smaller population subgroups. However, some analyses need a longer time frame; for these cases, a six-year file would be appropriate. Thus, the researcher can choose between a larger number of observations and a longer reference period.

The first available SLID data file will be the cross-sectional file for reference year 1993. The target date for release of this file is early 1995.

Longitudinal surveys are inherently difficult to process -- longitudinal consistency is not easily achieved. This is the primary reason for not proposing the release of longitudinal files each year. However, as many data users are already anticipating the availability of SLID longitudinal files, we are considering annual releases for the first panel only. Above and beyond the plans just outlined, this means that two-year, four-year and five-year longitudinal files will be available. (These exceptions are denoted by asterisks (*) in the table.)

Variety of Files

While this document refers to longitudinal files as if there is one type only, a range of products may be offered. For example, a summary longitudinal file, designed for easy tabulations, is a possibility. This file could emphasize derived variables that characterize changes experienced over time but suppress the details on spell length and timing. A product of this type might be a starting point for a researcher just beginning to use SLID data. A future SLID Research Paper will explore this possibility in more detail.

5. ORGANIZATION OF SLID LONGITUDINAL MICRODATA FILES

A working assumption of the SLID team was that each user would wish to create work files for analysis, containing only the variables and observations of interest for his or her project. Therefore, data-retrieval software will accompany all microdata files. In this way, data can be stored in the most efficient manner, but users will have a straightforward manner of selecting variables of interest as well as subsets of observations. This approach will also allow users to choose readily the most appropriate unit of observation: person, groups of persons living together (family, household), employers (SLID will be collecting some information from respondents in the context of each employer).

An issue yet to be decided is the output format created by the data-retrieval software. A rectangular format with a fixed record length, and each variable in a fixed position, could be the default format. This is a standard manner of presenting files generally, adaptable to most analytical software. In the longer term, it should be possible to directly create files in a format suitable for specific software. For example, two popular software packages for statistical analysis are SAS and SPSS. Both require data files in a format specific to their software. Therefore, it would seem useful to provide the option to directly create a SAS file or an SPSS file. Note that these are examples only, and user demand would determine the software packages for which direct file creation would be provided.

The output files will be provided on whichever medium is requested: CD-ROM, diskette, magnetic tape. It is possible that costs could differ for each medium, however.

6. STRUCTURE OF DATA VARIABLES

With data-retrieval software an explicit record layout is not needed, but users must still understand how the data are expressed. The basic guiding principle is to

present the data as if they had been collected once, at the end of the reference period covered by the file (for example, three years or six years). A conceptual (if not actual) process for achieving this objective follows.

SLID data are collected twice a year. The January interview examines labour market activities over the previous year. The May interview collects income data for the previous year. If one simply consolidated all data as collected, it would look like Figure 1. One observation would correspond to all data collected for a particular individual for a particular year.

Figure 1: Data Representation from Collection

PERSON 1 YEAR 1 DATA
PERSON 2 YEAR 1 DATA
•
•
•
PERSON N YEAR 1 DATA
PERSON 1 YEAR 2 DATA
PERSON 2 YEAR 2 DATA
•
•
•
PERSON N YEAR 2 DATA
•
•
•
PERSON N YEAR 6 DATA

Since longitudinal analysis is important, the data representation in Figure 1 would appear to be awkward to use for analyzing changes over time at an individual level. Therefore, the next step (again, in terms of a conceptual evolution) would be to place all data together for each individual, as illustrated in Figure 2.

Figure 2: Data Representation - One Observation per Individual

PERSON 1 YEAR 1 DATA ... YEAR 6 DATA
PERSON 2 YEAR 1 DATA ... YEAR 6 DATA
•
•
•
PERSON N YEAR 1 DATA ... YEAR 6 DATA

The next step in the evolution towards the objective is to rearrange the variables in each observation as represented in Figure 2. Figure 3 indicates this reordering to group data together for each variable.

Figure 3: Data Representation in Variable Order Rather than in Chronological Order

PERSON 1 (VAR 1 Y1) ... (VAR 1 Y6) ... (VAR M Y1) ... (VAR M Y6)
PERSON 2 (VAR 1 Y1) ... (VAR 1 Y6) ... (VAR M Y1) ... (VAR M Y6)
•
•
•
PERSON N (VAR 1 Y1) ... (VAR 1 Y6) ... (VAR M Y1) ... (VAR M Y6)

In many respects, there are no significant differences between the data representation given in Figures 1, 2, and 3, and particularly between the representations in Figures 2 and 3. They are simply a matter of different ordering. Certain redundancies would exist and could be removed. For example, date of birth need not be repeated for each year on the output file. The main issue is that ordering affects the degree of ease or difficulty a user faces in manipulating the file.

The next step is to remove "artificial data breaks" due to the way the data are collected. For example, SLID will be collecting the start and end dates of employment for each employer. Assume that a respondent worked for a particular employer from July 12 1993 to November 5 1995. The dates data as collected during the interviews are the following:

Reference Period	Start Date	End Date
1993	July 12	Ongoing at year end
1994	Continuing from previous year	Ongoing at year end
1995	Continuing from previous year	November 5

On a longitudinal file which includes reference years 1993, 1994, and 1995, the start and end date information for this employer can be more efficiently represented as Start Date = July 12 1993 and End Date = November 5 1995. No information is lost.

Fixed and Dynamic Variables

SLID variables will include those collected directly for a respondent and derived variables. Whether direct or derived, variables may be either "fixed" or "dynamic" in nature. A fixed variable cannot change over time, for example, date of birth. A dynamic variable is one which could change during the time a respondent is in the SLID sample, for example, marital status.

Note that the value of a dynamic variable may not change for every respondent. For example, the marital status of some SLID respondents will change during the survey reference period, and will not change for others.

For fixed variables, numerical values (usually) will be assigned to represent the various possible responses.

Values of dynamic variables will be represented in an identical manner -- the challenge is to efficiently, but simply, represent the changes in values. We are proposing the following:

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

where, T is the number of changes in values (defined separately for every dynamic variable for every respondent),

$X_1, \ldots, X_{T+1}$ are defined values of the variable,

D_i is the date of change from value X_i to value X_{i+1} ($i = 1, \ldots, T$).

The date of change will reflect the level of detail collected for the survey, which is a reflection of the anticipated accuracy of reporting. The possibilities are:

- Year / Month
- Year / Week
- Year / Month / Day

The second alternative (Year/Week) is a derivation from (Year/Month/Day). A concept which divides every year into 53 weeks will be used. A week is defined as a period from Sunday to Saturday inclusive. Week 1, which may be less than seven days, is the period from January 1 to the first Saturday in January. Similarly, Week 53, which may also be less than seven days, is the period from the last Sunday in December to December 31. With this definition, all years have exactly 53 weeks. The 53-week concept will be used for those variables defined for every week during the reference period and for those dates which are collected as Year/Month/Day, but for which it is felt that the reported day is a reliable estimate of a week during the month, but not of the exact day.

Modifications to this approach could be considered:

- Instead of numbering from 1 to 53 each year, consecutive numbering could be used. Thus, the first week of the second year would be week 54, and so on. In this scenario, the only difference is the numbering convention used.
- In keeping with the philosophy that the break between calendar years is not of primary importance, the same concept could be applied to the entire reference period. Thus, only the first and last weeks of the reference period could be less than seven days, and certain weeks would overlap calendar years. This approach would make the calculation of durations much simpler, as weeks would be numbered from 1 to 313 for a six year file.

If the last proposal is problematic due to weeks crossing calendar years, but durations are of great interest, a function for deriving durations could be supplied.

An example of the representation of a dynamic variable is given in Figure 4. The example uses the variable Labour Force Status, which can assume five values. Including the value 9 (Not in the SLID sample) illustrates another aspect of the data representation. SLID will collect information on "joiners"; i.e., all persons living with a SLID respondent, but who were not part of the household when the panel of respondents was introduced. An example is a person living with his/her parents when the household is selected as part of the SLID sample. Subsequently, this person moves out and gets married. The person's spouse is termed a "joiner".

Figure 4 - Example of Dynamic Variable Representation

Labour Force Status: Valid values are 1 = Employed / 2 = Unemployed / 3 = Not in the labour force / 8 = Unknown / 9 = Person not in SLID sample

Six-year longitudinal file covering reference years 1993-1998.

9	**9**	**4**	**0**	**1**	**1**	**9**	**6**	**3**	**1**	**2**	**9**	**8**	**1**	**7**	**3**	
X_1	Y_1		W_1		X_2	Y_2		W_2		X_3	Y_3		W_3		X_4	

Referring to the example illustrated in Figure 4, this person was a joiner, first picked up in SLID in January 1995. There is no information for reference year 1993. In the first week of 1994 (Y_1 = 94, W_1 = 01), the person was Employed (X_2 = 1). A subsequent interview determined that, in Week 31 of 1996 (Y_2 = 96, W_2 = 31), the person became Unemployed (X_3 = 2). Another change was noted in Week 17 of 1998 (Y_3 = 98, W_3 = 17), when the person left the labour force (X_4 = 3). No further changes are indicated, meaning that the person remained out of the labour force at the end of the survey reference period (in this example, December 31 1998).

While dynamic in nature, most monetary values will represented as:

X_1 X_2 X_3 X_4 X_5 X_6

This reflects the fact that no change dates are collected for these variables. In fact, for variables such as income sources, the reference period is the year, and no subannual information is collected.

7. CONTENT OF SLID LONGITUDINAL MICRODATA FILES

A detailed description of all variables is given in Part B of this report. In principle, all data collected for all respondents will be available to the data user through some mechanism which allows full access, but protects the confidentiality of individuals, as discussed in Section 3. Derived variables will be calculated and provided for "common" derivations.

The following information will be available from the file or retrieval software to allow "generalization" of user programs to deal with any SLID longitudinal files:

- Number of reference years covered by the file
- Start year of file reference period
- End year of file reference period
- Start and end dates for every week in all reference years
- Number of days in Weeks 1 and 53 for each of the reference years
- File count of all persons
- File count of persons aged 0-15, 16-69, 70 and over
- File count of person-employers (i.e., the sum over all persons of the number of employers during the reference period)
- File count of households, as per January 1 of each year
- File count of economic families, as per January 1 of each year
- File count of census families, as per January 1 of each year
- For each dynamic variable, the maximum number of changes recorded for any particular person (or employer, or household, or family, whichever is relevant) -- this corresponds to the maximum value of T as defined in the model representation of a dynamic variable.

The specifics provided in Part B describe the following groupings of variables:

- General person-level information
- Demographic Information
- Education -- attainment and activity
- Work History
- General labour information
- Employer-specific labour information
- Jobless spells
- Disability
- Support
- Income
- Wealth
- Information on other household members
- Household/family information

Information on survey content can also be obtained from SLID Research Papers 92-01A "Content of the Survey of Labour and Income Dynamics: Part A - Demographic and Labour Content" and 92-01B "Content of the Survey of Labour and Income Dynamics: Part B - Income and Wealth Content". Details on the actual questions, as developed for field testing in 1993, are available in SLID Research Papers 93-02 "SLID Labour Interview "Questionnaire" - January 1993" and 93-04 "SLID Income Interview - May 1993: Questionnaire and Data Collection Procedures".

As they fall outside the scope of both survey content and derived variables, two of the groups listed above are briefly explained here:

- General person-level information
    - Person identification code

- Geographic information on place of residence
- Date of preliminary interview (i.e., when person entered SLID sample)
- Sampling weights (to allow estimation of population counts from the sample)
- Proxy information (for each interview, who provided information, and if proxy, the relationship between the two)
- language of interview (for each one)

- Information on other household members

  This section provides information on all other persons living in the same household as the respondent, at any time during the reference period covered by the file. The following variables are provided for all such persons (some of these variables are dynamic):

  - Person identifier
  - Relationship between other person and respondent
  - Whether in same family or not
  - Date of birth
  - Sex

Survey data files contain "sampling weights" to be used to aggregate the individuals in the sample to the entire population. Weighting is more complex for longitudinal surveys than for cross-sectional surveys due to the fact that the population changes over time: people move in and out of the population (in terms of their place of residence), people are born and people die.

On longitudinal files, it is planned to include all persons for whom data were collected during the reference period (although perhaps not throughout the entire period). One longitudinal weight and one cross-sectional weight will be assigned

to each individual for each reference year. The weights for a particular year will be based on the population as of January 1 of the following year. The result is that certain weights for some respondents will be zero, if the respondent was not in the sample during a particular year. Thus, a basic understanding of weights will be required for data users to ensure that the proper weight is used for a particular analysis. More details on weighting will be available in a future SLID Research Paper.

8. DERIVED VARIABLES

In contrast to direct variables, derived variables are calculated from the information collected for a respondent. Proposed derived variables are listed in Part B of this report. Comments on their usefulness and suggestions for additional ones are welcome.

This section of the report deals with issues of a more general nature; i.e., those which are not specific to any particular derived variable.

The major issue deals with the calculation of derived variables for families or households when a change in the group composition occurs. For example, how should annual family income be calculated for families whose composition changes during the year?

- Family income could be calculated only for families whose composition did not change during the year; or,
- For each person, family income could be defined as the sum, over all persons in the same family at some point during the year, of income earned while a member of the family.

Clearly, the first alternative is simpler to calculate and to understand. It is also clear that the second alternative is a superior measure when conducting analyses at a person level where family income is an explanatory variable.

However, there are major difficulties in actually calculating the second measure, since no sub-annual income is directly measured. The survey will be collecting some information to help with the calculation, such as dates of employment, dates of receiving money from government programs such as Unemployment Insurance and Social Assistance, but no definitive sub-annual derivations are possible.

Another issue deals with the calculation of a derived variable in the presence of non-response. If a variable is derived from three others, and, for a particular person, one of the input variables is missing, how should the derived variable be calculated?

9. QUESTIONS TO PROSPECTIVE DATA USERS

While comments are welcome on any aspect of this report, a list of questions related to the key strategic issues is provided in this section, as a guide for those who wish to use it.

1. Does the proposed file availability meet your needs? Is your main interest in longitudinal files, cross-sectional files or both?

2. Are the "alternative" longitudinal files of (potential) interest? If so, what should the content be?

3. What is your reaction to the proposed representation of dynamic variables?

4. What is your reaction to proposed representation of dates, particularly the 53-week per year concept?

5. What is your reaction to proposed approach to provide data-retrieval software which would create work files for individual applications? What types of work files would you like it to provide?

6. Which software packages do you think you would use for analyzing SLID data? If there are more than one, is there one primary package?
7. What types of analyses do you expect to do with SLID data; for example, cross-tabulations, multiple regression, event history analysis?

8. What type of hardware would you have available for analyzing SLID data? If it is difficult to foresee what you might have in two or three years, what would you use if SLID data files were available now?

9. In what medium would you prefer to receive SLID data files? As with the previous question, if you cannot predict, what would be your preference now?

10. What is your reaction to the discussion on confidentiality and the data-access model? If you are in basic agreement, what types of test files or test data would be most useful to you?

11. Based on your anticipated data uses, do you have any suggestions on how best to calculate derived variables for families or households when a change in group composition occurs?

12. Is this approach to consultation useful? How could it be improved?

13. At what point would user workshops be of use? What should their content be? How should they be presented?

14. What should the relative priorities be for various SLID products?

15. Do you have any other comments which are not addressed by the other questions?