Catalogue No. 95-11

DISSEMINATING DATA FROM LONGITUDINAL SURVEYS: ISSUES FACING THE SURVEY OF LABOUR AND INCOME DYNAMICS

Product Registration Number 75F0002M

May 1995

Maryanne Webber, 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. 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 INTERNET (GILES@STATCAN.CA), by telephone (613) 951-2891, or by fax (613) 951-3253.

EXECUTIVE SUMMARY

This paper was presented at the CAPDU-IASSIST joint session in Québec City (May 1995).

The dissemination of microdata from longitudinal surveys poses several challenges. The purpose of this paper is to outline these challenges and some of the measures being proposed to deal with them. The paper begins with a brief overview of the survey content and design as context, but the main purpose of the paper is to provoke discussion on general dissemination issues, using SLID as a case study. The intended audience is research librarians and others who will play a role in the dissemination process.

TABLE OF CONTENTS

1. INTRODUCTION

The Survey of Labour and Income Dynamics (SLID) is one of several new longitudinal household surveys being mounted by Statistics Canada. Like the others, SLID is preparing for the release of its first round of microdata. The dissemination of microdata from longitudinal surveys poses several challenges. The purpose of this paper is to outline these challenges and some of the measures being proposed to deal with them. The paper begins with a brief overview of the survey content and design as context, but the main purpose of the paper is to provoke discussion on general dissemination issues, using SLID as a case study. The intended audience is research librarians and others who will play a role in the dissemination process.

2. OUTLINE OF THE SURVEY

SLID is designed to track the experiences of individuals in the labour market, their level and sources of income and changes in family life over a period of six years. The first panel began in 1993, with labour and income information collected from about 31,000 persons aged 16 and over. A second panel will begin in 1996, doubling the sample size. In 1999, when the first panel ends, a third one will begin. This approach of rotating, overlapping panels ensures that the sample remains representative.

During the six years, 13 interviews are conducted. A preliminary interview is done when a panel first starts up, to collect background demographic, education and work experience information. One year later, an annual cycle of labour and income interviews begins. Every January, information on the person's labour market activities throughout the previous year is recorded; in May, income sources and amounts for the previous year are collected.

A summary list of variables from the survey and a chart depicting the main types of information are presented in appendix. Major research areas will range from employment and unemployment dynamics and labour market transitions linked to the life cycle, to job quality, workplace inequality issues, family economic mobility (dealing with shifts in income level), low income dynamics (or flows into and out of poverty), demographic events and the relationship between work and education. Researchers are expected to come from many disciplines.

3. DATABASE SIZE AND COMPLEXITY: THE MAIN CHALLENGE

By household survey standards, the SLID database will be large and complex. Even with our best efforts to make it approachable, researchers will need to make an "up front" investment of time and effort to come to grips with it. Why is this so?

Number of variables and hierarchical structure

Perhaps the most fundamental reason is the size of the dataset and its internal relationships. As a rough estimate, there are 500 distinct variables in the full dataset, without taking the time dimension into account. This means that events, spells, variables collected annually and variables collected as many times as applicable are all counted only once -- and there are many such variables in the dataset.

Hierarchical relationships abound in the data. A person can have several employers and information is collected on up to six jobs per year. There may be several work absences from each job. Over time, even if a person does not change employers, he or she can have several occupations, wage rates and work schedules. The survey will also yield information at the household and family

level. Because of the hierarchical nature of the survey content, we are processing the data in a relational database environment and are also proposing to use a relational database for the microdata output.[1]

Time dimension

Like all longitudinal surveys, SLID users will need to grapple with the time dimension. From the time perspective, we can distinguish different types of variables. First, variables like gender, year of birth and ethnic origin, are *fixed*. If an error is detected these variables may be corrected but otherwise they do not change over time. Next, there are *annual* variables, such as weeks worked during the year and investment income. For these variables, the reference period is by definition the calendar year. Thus, for a full panel, there will be six observations for each record. There are also *cumulative* variables, like years of schooling, years of work experience and number of children where, depending on the respondent's activities or circumstances, the values may or may not require updating each year. Finally there are *dynamic* variables which relate to spells. The duration of a spell may range from a week to several years. SLID's content includes many variables expressed as spells and, to facilitate analysis, spells that cross the seam between two reference years (for example, an unemployment spell that begins in November and ends the following March) will be linked up on the database. In effect, the dataset will ultimately look as if the information was collected retrospectively at the end of the six years, as opposed to being a series of unrelated snapshots.

---

1 The first wave (including results from the preliminary interview) will, however, be released as a rectangular file. The content has not yet been finalized but our best estimate is that the record length will be about 3000 bytes for a total file length of roughly 90 MB. Every year, the dataset grows, i.e., the second year's file will incorporate and replace the first.

Units of analysis

Another factor that adds to the learning curve -- and this again is due to the hierarchical properties of the data -- is that there are many possible units of analysis. The *person* is the basic unit. In addition to being the appropriate unit for many types of research focused on the individual, the person will also generally be used for studies of the family. Because family composition can change over time, the definition of family poses some sticky problems in longitudinal research. One can however define the person as the unit of analysis and develop typologies to characterize the person's family circumstances over the study period.

The *person-job* is a unit of analysis used with data from labour market surveys with a one-year reference period, like the Survey of Work History and the Labour Market Activity Survey. We expect that researchers will also use the person-job for SLID studies. This unit of analysis came about as a way of handling the fact that a person may have several jobs, concurrently or consecutively, during a one-year period. Instead of using complex and arbitrary assumptions to select a main job for the year, all jobs are included and weighted using the respondent's sample weight. Sometimes they are further weighted by annual hours worked, so that part-time jobs lasting one month are given less weight than full-year, full-time jobs.

Some studies will use *spells* as the unit of analysis. For example, if a person is unemployed for two separate stretches during the study period, the two spells of unemployment will be included, both receiving the respondent's sample weight. Demographic and other characteristics can be treated as attributes of the spell. Similarly, researchers may use *transitions* as a unit of analysis. Some transitions can be identified from dynamic variables, when one state ends and another begins. Some data users will no doubt want to develop definitions of transitions tailored to a particular study. For example, it should be possible to use SLID to study work-

to-retirement transitions or job promotions. But since these are complex processes, there is no variable or flag on the database identifying these events. Rather, the user will need to look at a range of variables and explicitly define the event of interest.

4. TOOLS TO HELP RESEARCHERS GET STARTED

The survey staff are very aware of the challenge data users face in getting started. It is incumbent on us to develop tools and user support strategies that increase data accessibility. What are these tools and strategies?

Database design

Because of the size and complexity of the data, a data model was developed. This is a device for structuring the survey content and giving explicit expression to the relationships in the data. The development of the data model was done following two important principles, both of which were intended to aid the data user.

First, variables were defined in keeping with the survey's content objectives, rather than as a simple reflection of the questions and response categories used in data collection. The survey questions are designed to accommodate data collection, and are often not that useful as analytical variables. For example, to collect one content item, there may be several different questions addressed to various subgroups.

Second, the decision to collect data annually was based on respondent recall and other operational considerations. It was decided that this feature of the data collection operation should be transparent in the output variables (except of course in cases where annual observations make sense from a content point of view). The

data for a six-year panel should look like they were collected once covering the full six-year period.

These principles required a significant “up front” design and development effort but hopefully they will pay off in downstream benefits to data users who would otherwise have to recreate “seamless” data from a series of snapshots.

Software to retrieve data from database

We are planning to provide a public-use microdata file with front-end software that, at a minimum, allows users to select variables and subpopulations of interest, for specified time frames. These smaller datasets can then be downloaded into a flat file for further analysis using whatever software the user chooses. There will also be easy ways of producing simple frequency counts from the full dataset, to help users define their study populations.

CD-ROM

The public-use microdata file will be available on a CD-ROM. This will hopefully increase data accessibility.

Major reference products

There are three types of documentation in the works: technical documentation of the database content and structure; a user handbook; and research papers providing detailed documentation on specific topics.

The main SLID database is being designed with the technical user documentation -- variable names, descriptions, definitions, algorithms for derived variables, code

lists and user notes -- as an integral part. This documentation is being stored in a relational format, so it is possible to extract parts and produce customized reports. Microdata users will be able to access the documentation electronically as it will be imbedded in the product.

A handbook or "friendly" user guide is also being developed. This should be of interest to users of custom tabulations as well as to actual and potential microdata users. After the first few editions, this publication will probably stabilize and enjoy a relatively long shelf-life -- perhaps we will re-issue it every six years to coincide with the completion of a panel.

Finally, SLID has a general purpose research paper series. Since 1992, we have produced about 15-20 of these reports each year. We are beginning to use this series as a repository for detailed information on specific variables, for example, the composition of "roll-up" categories for mother tongue and ethnic origin.

Workshops

To get started, some users may be interested in participating in a workshop. We are quite sure that there will be interest in a workshop on the content and structure of the database. We have already been asked by a few groups to do workshops of this type and have agreed. There may also be interest in analytical techniques appropriate for use with these data.

Sharing information on research in progress

Throughout the survey development process, decisions and issues have been documented in the quarterly newsletter, *Dynamics*. While there will still be developments to communicate in coming years, we expect that the role and

content of *Dynamics* will gradually shift, hopefully becoming a forum for exchange on research underway outside as well as inside Statistics Canada. It is very beneficial for the survey staff and the Agency to be aware of data uses (as well as research *not* being done because of the lack of a few key variables). Short research summaries in *Dynamics* would keep us up to date and could supplement whatever other exchange mechanisms exist among researchers in a particular field.

5. CONFIDENTIALITY

Longitudinal surveys in general face a challenge because the events and transitions that they document -- and that are central to their analytical potential -- may create risks of disclosing the identity of respondents. Moreover, when the first wave is released, it is impossible know what patterns of change over time will be common or rare several years down the road, which means that we may need to reconsider the content of the public-use file as the data from successive waves build up.

In SLID's case, there are difficult trade-offs between geography, family information and labour market detail. The data are supposed to meet the needs of researchers in a range of disciplines and to allow analysis of the interactions that exist between labour market behaviour, family circumstances and income. This makes it very difficult to protect confidentiality without "short-changing" any particular user group.

The search for solutions is very lively. Research is under way on techniques for quantitatively assessing disclosure risk and on alternatives to suppression and collapsing. Other statistical agencies are being consulted on their approaches. An attempt is being made to prototype a remote access system, which would allow researchers to write and test their programs off-site and telecommunicate them to us so we could execute them against the full database. We are also investigating

the possibility of licensing researchers to use a middle-level file for a specified purpose, following stringent rules regarding access, security and disposal. There is enough concern and energy being devoted to this issue to hope that solutions will emerge.

In the meantime, we are defining the content of a public-use microdata file that would be screened using the usual Statistics Canada procedures. Several analytically interesting derived variables are being added to the file to reduce the impact of missing detail. Here are a few examples:

- several occupation typologies;
- the relevant low-income cutoff, or a measure showing family income as a ratio of the relevant LICO;
- a derived variable showing the link between occupation and major field of study.

Hopefully, variables such as these will help researchers to proceed with their work even if some of the very detailed information (like 4-digit occupation) is not on the public-use file.

We also face a dilemma with respect to family information. On the main base, it is possible to link up family members (and previous family members) but, to provide this capacity on the public-use file, it would be necessary to reduce the amount of labour market information. As a compromise, we are proposing to include a good range of family variables, but only for a subsample of respondents. This means that researchers have access to more variables on the public-use file and, should they require results for the full population, the same program can be re-run against the full data base. These measures will ensure that, even if some variables are missing, the public-use file will still be a rich source of information.

6. COMPUTER-ASSISTED INTERVIEWING AND USER DOCUMENTATION

Although it does not exclusively concern longitudinal surveys, the move to computer-assisted interviewing for household surveys at Statistics Canada is raising some interesting documentation issues. We are finding that efforts to document the questionnaire are proving to be very labour-intensive and error-prone. We have been searching for tools and techniques to improve the process and trying to promote some measure of consistency across surveys.

A working group was set up recently in the household surveys area to address this issue. It looked at a number of options. One idea was to produce a print image of each screen. However, this would yield very bulky documents and, for surveys with complex branching (like SLID), it would be nightmarish to follow flows. Also, even with that level of detail, many special features such as hot keys and edits would not automatically be documented. Similarly, the idea of producing a diskette with the questionnaire is appealing at first blush but this would not be very meaningful as a "stand-alone" product. The user would need to learn the data collection software. Moreover, many survey applications -- particularly longitudinal ones -- do not start with a blank sheet. There are prefilled items that affect questionnaire flow. Without these prefills, one cannot get into various branches of the application.

After examining these and other options, the working group found that, at least for the time being, the best approach is to concentrate on producing a good survey codebook. Among other advantages, this is an approach where standards or guidelines across surveys are a reasonable goal and where the documentation reflects the data user's perspective. This means that the user documentation of a questionnaire would begin with the output variables and work backwards, ending

with the questions underlying the variables. Instead of expecting users to follow complex flows through hundreds of questions, each question or group of questions would have a "universe statement" describing the question's target population.

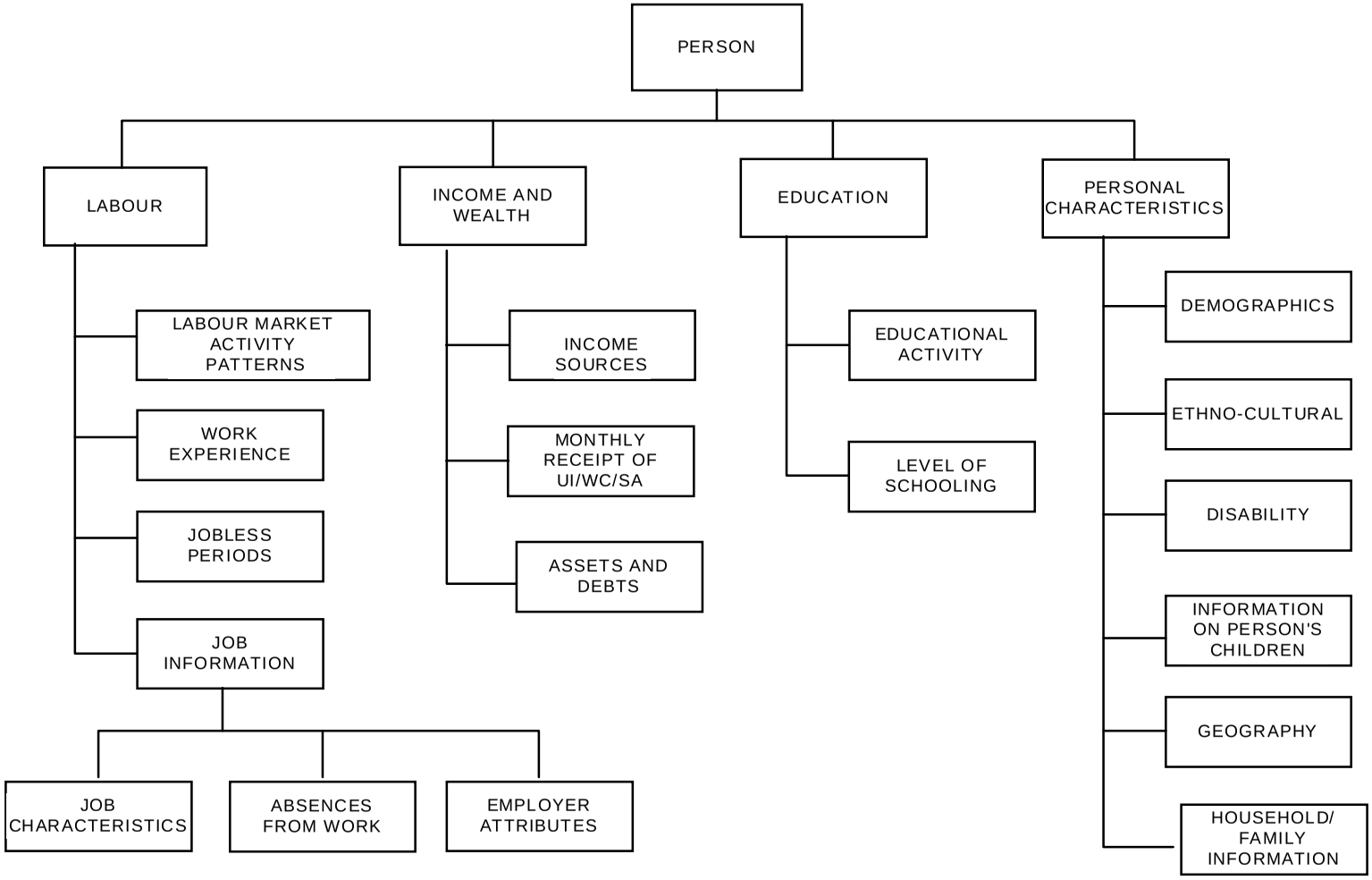
The group also concluded that different surveys would require different supplementary tools, depending on audience, length, complexity and periodicity. In SLID's case, flow diagrams showing the organization of the survey content at increasingly detailed levels are being developed.

7. CONCLUSION

Once established, longitudinal surveys can be invaluable -- but it can take time to become established. In the current fiscal and social policy climate, time is at a premium. New longitudinal surveys cannot afford many years to demonstrate their value. There is therefore a pressing need to support researchers in getting started. In this paper, some of the dissemination measures planned for SLID have been reviewed. Feedback on current plans will help us to get off to a good start. At the same time, this is a learning experience for survey staff as well as researchers. We fully expect to make adjustments to products and services and therefore hope to sustain a dialogue on enhancements.

APPENDIX: ORGANIZATION OF SLID CONTENT AND PARTIAL LIST OF VARIABLES

SURVEY OF LABOUR AND INCOME DYNAMICS: ORGANIZATION OF CONTENT



Partial List of Variables

I. Labour

Nature and pattern of labour market activities

- spells of employment and unemployment (start and end dates, durations)
- weekly labour force status
- total weeks of employment, unemployment and inactivity by year
- multiple job-holding spells
- work absence spells

Work experience

- years of full-time and part-time employment
- years of experience in full-time, full-year equivalent

Characteristics of jobless spells

- job search during spell
- dates of search spells
- desire for employment
- reason for not looking

Job characteristics (all characteristics updated each year and dates of changes recorded; collected for up to six jobs per year)

- wage
- work schedule (hours and type)
- benefits
- union membership
- occupation
- supervisory and managerial responsibilities

- class of worker
- tenure
- first date ever worker for this employer
- how job was obtained
- reason for job separation

Characteristics of work absences lasting one or more weeks (collected on first and last absence each year, for each employer)

- absence dates
- reason
- paid or unpaid

Employer attributes

- industry
- firm size

II. Income and wealth

Personal income

- annual information on about 25 income sources
- total income
- taxes paid
- after tax income

Receipt of compensation (whether benefits were received from each source and, if so, in which months)

- Unemployment Insurance
- Social Assistance
- Worker's Compensation

Assets and debts

Information might be collected once or twice in life of panel on roughly 20 asset and debt categories.

III. Education

Educational activity

- enrolled in a credit program, months attended
- type of institution
- full-time or part-time student
- certificates received

Educational attainment (updated annually)

- years of schooling
- degrees and diplomas
- major field of study

IV. Personal characteristics

Demographics

- year of birth / age
- sex
- current marital state and date it began
- year/age at first marriage

Ethno-cultural

- ethnic background
- member of an Employment Equity designated group
- mother tongue
- date of immigration
- country of birth
- parents' schooling

Activity limitation

- annual information on activity limitations and their impact on working
- satisfaction with work

Information on person's children

- number of children born, raised
- year and person's age when first child born

Geography and geographic mobility

- economic region or CMA of current residence
- size of community
- moved during year
- move dates

- reason for move
- nature of move (full household/household split)

Household and economic family information (annual summary information, e.g., size, type)

- key characteristics of other individuals in household (e.g., age, sex, relationship, income, annual hours worked)
- household/family size and type
- family income
- relevant low-income cutoff
- family events (separation, death, birth)
- dwelling type and tenure