Catalogue No. 95-16

MODELLING DATA FILES FOR LONGITUDINAL SURVEYS

Product Registration Number 75F0002M

August 1995

Philip Giles, Household Surveys Division

Pierre Lafrance, 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 1995 Annual Meetings of the American Statistical Association in Orlando, Florida.

Traditional survey data files have been developed by providing the responses to each of the survey questions in the order listed on the questionnaire. A series of derived variables (i.e., derived from two or more of the survey responses) are added to the end of each record. While still feasible, recent survey developments, namely the widespread introduction of longitudinal surveys and computer-assisted interviewing, have greatly complicated this approach.

This paper examines the development of survey data files, or data modelling, for longitudinal surveys.

TABLE OF CONTENTS

1. Introduction

Traditional survey data files have been developed by providing the responses to each of the survey questions in the order listed on the questionnaire. A series of derived variables (i.e., derived from two or more of the survey responses) are added to the end of each record. While still feasible, recent survey developments, namely the widespread introduction of longitudinal surveys and computer-assisted interviewing, have greatly complicated this approach.

This paper examines the development of survey data files, or data modelling, for longitudinal surveys. Following a general description, the paper examines a particular new longitudinal survey, the Survey of Labour and Income Dynamics conducted by Statistics Canada.

2. Key elements of a longitudinal survey

With recurring surveys, it makes a difference whether you interview a new sample of people each time, as most surveys do, or the same people several times in a row, as in a longitudinal survey. The reputed advantage of cross-sectional samples (i.e., sample for each collection is different, at least partially) is that they are generally more representative of the population, and they more accurately reveal the levels and trends for the whole population and for various subgroups.

In longitudinal surveys, the focus shifts from static measures and net change to the whole range of transitions, durations, and repeat occurrences of the characteristics measured by the survey. Many longitudinal surveys use a recall period equal to the time between interviews. For example, a longitudinal survey which collects data once a year asks the respondent to report information covering the previous twelve months. In this way, at least conceptually, there are no gaps in time for the information pertaining to every respondent. [1]

Another important aspect of longitudinal surveys is the development of “following rules”, the set of procedures for deciding who to interview for each data collection. Often, these are simple in principle, but operationally very complex. For example, most longitudinal surveys wish to follow persons when they change residences. But several types of moves are possible. For some, the move may involve a change in circumstances such that the survey content is no longer appropriate. People may move out of the country or be institutionalized. It may be desirable to follow these people to track changes in residence, but not to interview them for the survey.

Just as the development of longitudinal surveys raises new issues for data collection, the development of data files poses new challenges. This document aims to identify these challenges and to discuss how they are being met by a particular survey.

3. Approaches to data modelling

First, some basic definitions are needed. (Although some of the definitions and principles described here are more widely applicable, the discussion will focus on survey data files.) In very simplistic terms, data modelling is the development of a database structure to store the survey data. A more detailed explanation is given

[1] Standard survey problems such as nonresponse and recall errors render this “continuity in time” concept only partially true. As well, longitudinal surveys which do not collect data “continuously” are unable to determine lengths of spells (for example, marital spells or spells of unemployment) or to sequence events in all cases. (Which came first: the person losing his job or separating from his wife?) However, these surveys may not have these objectives, so the loss of this information is not important.

later in this section. A database (DB) is a collection of data, organized to enable easy recording, retrieval and maintenance of the information. A Database Management System (DBMS) is a collection of software programs used to manage a database. All requests or programs for adding, manipulating or retrieving data from the database are handled by the DBMS.

Traditionally, survey databases were “flat” files, with one record per survey unit (for example, person or business establishment). All data variables relating to the survey unit were stored on the same record. Surveys are becoming more complex; longitudinal surveys being just an example. The traditional approach for data files has some disadvantages, so other types of databases are being considered. Three such types are:

- the relational approach
- the hierarchical approach
- the network approach

A description and comparison of these three is beyond the scope of this paper. However, of these three approaches, the relational approach is clearly becoming the standard in the industry. Oracle, Paradox, FoxPro, DB2 are examples of Relational Database Management Systems (RDBMSs) being used. Thus, discussion will be limited to relational databases.

3.1 Relational Databases

A relational database is a database that can be viewed as a collection of ***relations*** or two-dimensional ***tables***. The following is an example of a sample relational table; call it PERSON:

PERSONID	NAME	AGE	SEX
011100000001	Luc	22	M
011200000002	Paul	37	M
021100000003	Diane	27	F

The terminology is drawn from mathematics. Each table is a special case of the mathematical construct known in mathematics as a ***relation***. A row is analogous to a ***tuple***, and a column is an ***attribute***. To non-systems analysts a relational table may be viewed as a flat file. Strictly speaking, a relational table has a much more narrow definition than a flat file. But, for the purposes of this paper, they are considered as equivalents.

The relational database may be comprised of more than one table. In fact, for this paper, it is this possibility which distinguishes a relational database from a flat file. Consider another table which lists the jobs a person has; call it JOB.

PERSONID	JOB#	HOURLY WAGE	YEARS WORKED
011100000001	1	$ 7.50	1.50
011100000001	2	$ 5.35	1.00
011200000002	1	$ 14.75	12.75
021100000003	1	$ 10.00	7.00
021100000003	2	$ 9.25	1.50
021100000003	3	$ 12.30	0.25

By looking only at table JOB, one can see (by looking at records with the same PERSONID) that there are three persons represented, the first has two jobs, the second one, and the third has three jobs. However, the values of PERSONID are the same ones used in table PERSON. By linking the two tables, with records matched using the value of PERSONID, one can obtain more information, such as the age and sex of the person holding the job.

In terms of commands to manipulate relational databases, the Structured Query Language (SQL) is an ANSI (American National Standards Institute) standard. By using software which adheres to this standard in the creation of relational databases, one can easily use other database software when accessing the data.

Several advantages result from the use of a relational database.

- Relational database tables are simple but disciplined.
- Data that are logically related can be stored in one table (ex., information related to one certificate or diploma obtained by one person). A user can then focus on the areas of the database that are of interest to him/her.
- Redundancies in the data can be reduced. (In the above example, the person's age and sex need not be stored for every job.)
- Less space requirement to store the data in a relational structure.
- The data can be shared.
- Inconsistencies in the data can be avoided, and data integrity can be maintained.
- Standards can be enforced (ex., all dates have a common format).
- Common source of information controlled by a central staff.
- Security restrictions can be applied, different views on the data can be created (ex., grant access to certain tables or certain fields within a table).

3.2 Describing the Data Model

In informatics terms, a data model is a graphical representation of real life objects or entities and how they relate to one another. In the example used above, these objects are a person and a job. Data modelling consists of deciding exactly what information is to be held in the database. It is to identify the entities of interest and the information to be recorded about these entities. Data modelling also consists of identifying the relationships between the entities.

Data modelling is done through the use of Entity-Relationship Diagrams (ERDs) and the Data Dictionary (DD). Examples of each, taken from the Survey of Labour and Income Dynamics, are given in Appendix 1.

The ERD declares the components and the connections, and has the following basic elements:

C Entities

These correspond to the tables. In the example, these would be PERSON and JOB.

C Relationships

Entities relate to one another, or in other words, there exists relationships between these entities, and the cardinality of the relationships can be specified. In the example, each Person ***may work at*** 1 or many Jobs.

C Attributes

Entities have attributes that describe them. In the example, *Age*, *Sex*, and *Name* are all attributes of the Person entity. The attributes can be shown on the ERD, or if they are too numerous, the attributes may be described in the Data Dictionary.

The Data Dictionary (DD) accompanies the ERDs and defines the components. Each entity should have a data dictionary entry containing:

C a definition of the purpose or the role played by the entity in the model

C the data elements or attributes which describe the entity

C identifiers which uniquely identify instances of the entity

Performing the conceptual analysis to define the data model requires a lot of effort and time, but it speeds up the implementation phase. Once one creates a data file corresponding to the model, the elements of the data model have equivalent elements in the data file:

Data Model	Data File
Entities	Tables
Attributes	Fields
Identifiers	Key Fields

The following advantages stem from the development of a data model:

! Graphic ER diagrams (see appendix 1 for an example) are easy to understand not only for systems analysts but also for managers, subject matter specialists, methodologists and users.

! The ERD and DD provide an excellent basis for good documentation.

! Both are good communication tools between the systems people and the subject matter specialists during the process of identifying user information requirements.

! The development of the data model forces you, at an early stage of survey development, to understand what your data is or what you want your data to be in very specific detail. It clarifies the conceptual framework, avoiding the preclusion of any analytical possibilities.

! A data model provides an organization structure for the data.

! If the representation of entities and how they are related is valid in the model (i.e., in theory) it would also be valid in the creation of the data files (eg., relational database).

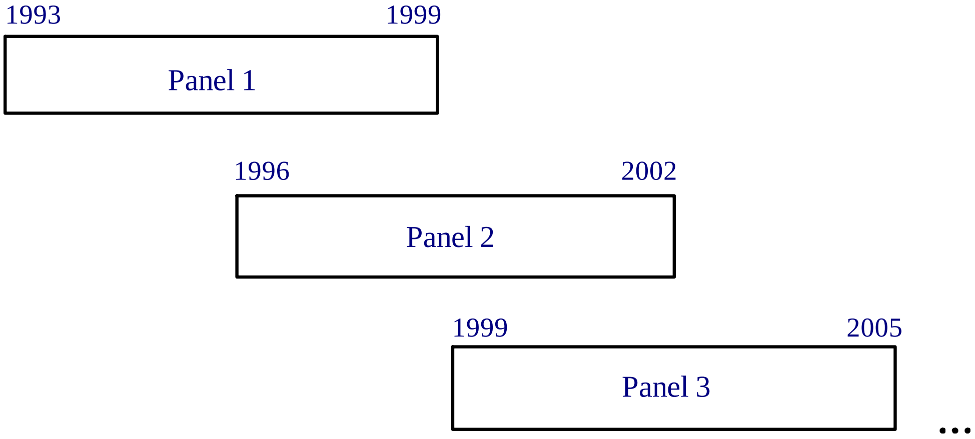
! A data model simplifies the task of developing and implementing a processing and production system.

! When developing a processing model and system, the existence of a data model will simplify the identification of processes that can/cannot be performed. It provides an opportunity to envision all the useful attributes that can be derived (by calculation, etc.) from the survey answers, and to define them and their relationships from the outset.

4. Overview of the Survey of Labour and Income Dynamics (SLID)

The Survey of Labour and Income Dynamics (SLID) is one of several new longitudinal household surveys being mounted by Statistics Canada. The survey 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.

Figure 1: Overlapping Design of SLID Panels



The first panel of respondents began in 1993, with labour and income information collected from about 31,000 persons aged 16 and over (in about 15,000 households). 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. There will be six years of longitudinal data for about 31,000 persons and three years of longitudinal data for double this number from a common set of respondents from two panels. Figure 1 illustrates the sample rotation.

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 retrospective 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. All survey data (with the exception of the preliminary interview in January 1993 with the Panel 1 respondents) are collected using computer-assisted interviewing (CAI).

All household members who were in the selected dwellings in January 1993 are considered longitudinal respondents and will be followed for six years— until the end of 1998. As well, people who move in with longitudinal respondents during the six years are also included and will answer the same questions as longitudinal respondents. Over time, the number of households in the panel will grow as household splits occur. This will be offset by reductions due to attrition.

A summary list of variables from the survey and a chart depicting the main types of information are presented in Appendix 2. A review of these variables will reveal the complex data relationships in the SLID variables, giving a direct indication that data structures to store these variables will also be complex.

5. The SLID data model: Guiding principles

Several early decisions were important to the development of the SLID data model. These became "guiding principles" throughout development.

Take a long term view to development, rather than to start with immediate needs and modify as required over time.
This was the first major decision related to the output files. This implied a longer development time for the first output files, but this was deemed to be more than compensated by other factors. First, data users would not have to adjust to a changing product over time. Second, it was believed that following this approach would result in future file releases (by the third, fourth or fifth year) being released earlier than would otherwise be the case. Third, after an initial period of familiarization, the approach would be clearer to the SLID project team, particularly those involved in processing. It would eliminate a "moving target".

Design the file structure to hold maximum data (i.e., six years) rather than to build from year to year.
This was an immediate application of the first principle, and could even be considered to be part of it. It meant that there was a need to think through what the file should look like six years from now, a difficult but useful exercise.

Output variables should be meaningful to data users, not just a reflection of collection questions (i.e., files should be variable based and not question based)
This explicitly recognizes that there is not a one-to-one correspondence between what is analytically useful (variable) and a question in data collection. The vast majority of users are interested in the variables and not necessarily how they were collected.

This approach lengthened the time required for the development of the data model and is increasing the development time for processing systems. However, we felt that this approach would help simplify a complex data file by reducing the amount of data manipulation required by data users. A concern of ours is that those who are less technically able will not use SLID data files due to its complexity. This is just one measure being taken to deal with this concern.

Output data should look as if it were collected once after six years

We are collecting data annually. Thus at the end of a panel, we have six sets of data for each respondent, each set covering a different one year period. The most straightforward approach is to provide the user with data in this form. However, in an analogous manner to our approach to data variables, we will merge the data across years to reduce data manipulation required by data users.

Figure 2: Job spells crossing "seams"

	Prior to survey						J	F	M	A	M	J	J	A	S	O	N	D	Second reference year								
Job 1		**M**	)	)	)	)	)	)	**M**																		
Job 2								**M**	)	)	)	)	)	)	)	**M**											
Job 3											**M**	)	)	)	)	)	)	)	)	)	)	**M**					
Job 4													**M**	)	)	)	)	**M**									
Job 5				**M**	)	)	)	)	)	)	)	)	)	)	)	)	)	)	)	)	)	)	)	)	)	)	>
Job 6								**M**	)	)	)	)	)	)	)	)	)	)	)	)	)	)	)	)	)	)	>

A simple example, illustrated in Figure 2, can illustrate how we are eliminating "seams" in the files.

The example shows six jobs which existed during the first reference year of the survey.

C Job 1 starts before the survey begins and ends during this first reference year. The survey collects the start date of the job, but all other data for the job pertains to one year.

C Job 2 starts and ends during the first reference year. Again, the survey collects data for only one year.

C Job 3 starts during the first reference year, but ends during the second reference year. Data on this job are collected twice, once for each year, as some characteristics can change over time. However, the structure of the data file will put the two pieces together rather than force the user to match data from the two years.

C At the end of the first reference year, job 4 will look like job 3 since both existed on December 31. However, after the second reference year, they will look different. Without the elimination of the seam, it would be necessary to look for a match in year 2 for job 4.

C Jobs 5 and 6 continue for several years. The seams are eliminated for all years.

The structure of public use microdata files is identical to that of the internal master file.

This principle does not have as significant an impact as the others in terms of the development of the data model, but has important ramifications on simplifying user access to SLID data. To preserve confidentiality of respondents, two types of data manipulation are usually performed on the internal data file: all values of some variables are suppressed (i.e., the variable is deleted; for example, date of birth is not provided) and values of some variables are collapsed (for example, providing age ranges rather than individual years of age).

The structure of SLID public use files will contain the suppressed variables. However, no data will be provided on the public use data file for these variables. Variables with collapsed values will be treated as suppressed; that is, the variable will be in the data structure but no data will be included. The collapsed version of the variable will be added to the data model, with data provided on the public use microdata file.

Why would one provide data variables with no data? Currently, data users wishing to use survey data on internal Statistics Canada data files must request the information from the survey staff. Provided that the information request does not breech confidentiality restrictions, it is produced by the survey staff and sent to the data user. The data user must pay all charges incurred in satisfying the request, such as person time and computer time. As well, depending on the complexity of the request, there may be a time delay in providing the information.

The theory behind providing an identical structure on public use microdata files is that users are able to see the full range of variables available and to write a computer program themselves to extract the required information. The program can be transmitted electronically to Statistics Canada. A member of the survey staff can run the program and review it for confidentiality considerations, and transmit the results electronically to the data user. This approach will reduce the time involvement of Statistics Canada staff, which will lower the cost of data requests. It should also reduce the turnaround time. This so-called “remote access” approach is not currently in place at Statistics Canada, but development is under way, and SLID wishes to be ready to use it.

6. Developing the SLID model

The following steps were followed to develop the SLID data model. A CASE (Computer Assisted Software Engineering) tool was used to electronically record the development. The CASE tool stored information in a relational database which provided an opportunity for using it in the development of the data file codebook.

C The starting point was the definition of the survey content themes. Appendix 2 outlines the SLID content themes; each box in the diagram represents a theme.

C The survey questions were linked to the themes.

C The element of time was defined for the questions. That is, some questions are asked only once as they are not inherently subject to change (for example, date of birth). Other questions are asked annually, but no date of change was collected. Other questions were asked annually and dates of change were recorded, thus providing a continuous history for that question.

C The various units of collected data were identified. Examples of SLID units are person, job, jobless spell, university degree.

C The list of entities was created. An entity was created for different themes, different time dimensions and different units.

C The questions were assigned to entities, from which variables were defined.

C Derived variables (those which can be defined from one or more variables) were defined. These can be variables with collapsed detail or summaries of several variables. As SLID data are collected separately for individuals, all family level variables are derived.

7. Issues in the development of the SLID data model

Longitudinal vs. cross-sectional data uses

While the main objectives of SLID are to allow longitudinal analyses, the SLID dataset is very comprehensive and unique even at a cross-sectional level (i.e., for a particular year). The design of the SLID data model should allow cross-sectional use.

Units of analysis

SLID data are collected for each person in the sample. However, many variables pertain to a unit of analysis which is not the person. For example, job characteristics pertain to a particular job which the person holds. However, at any time a person may hold no job, one job, or more than one job. Other units of analysis in SLID data are: absences from a job, household and family (discussed in more detail below), and a lengthy list of possible spells, such as marital spells, jobless spells, unemployment spells.

Representing time: Spells vs vectors

A major consideration for a longitudinal survey is how time is represented on data files. There are two primary approaches, the spell approach and the vector approach, each being useful for different types of analyses. The spell approach is to link every value for a variable to a start and end date, whereas the vector approach represents continuous time as discrete values. The periodicity of the vector components is dependent on the frequency with which the value can change.

An example is presented to illustrate the two approaches. Consider the labour force status of a particular individual over a given year. Assume that this person was employed from January 1 to March 11, unemployed from March 12 to May 7, employed from May 8 to May 26, out of the labour force from May 27 to September 10, and unemployed from September 11 to December 31.

According to the spell approach, this person would have 5 spells for the year as follows:

C Employed from January 1 to March 11;
C Unemployed from March 12 to May 7;
C Employed from May 8 to May 26;
C Out of the labour force from May 27 to September 10;
C Unemployed from September 11 to December 31.

By combining all spells for the year, one has a continuous description of the person's labour force status for the year. This provides complete information, but the number of spells would vary from person to person, complicating the use of the data. Two possible monthly vectors could be created, one which represents the situation at a given time in the month and the other which combines all values for a given month. Assuming that we are interested in the value at the end of each month, the following table provides the two vectors:

Month	**End of month value**	**All possible values**
January	Employed	Employed only
February	Employed	Employed only
March	Unemployed	Employed and unemployed
April	Unemployed	Unemployed only

May	Out of labour force	Employed, unemployed, and out of labour force
June	Out of labour force	Out of labour force only
July	Out of labour force	Out of labour force only
August	Out of labour force	Out of labour force only
September	Unemployed	Unemployed and out of labour force
October	Unemployed	Unemployed only
November	Unemployed	Unemployed only
December	Unemployed	Unemployed only

One can see the advantages and disadvantages. The vector which identifies a value at one point in the month is far simpler, but does not provide as much information as the spell representation. In fact, it misses the short employment spell in May. The second vector represents a compromise between the two. It is more complicated than the first vector since each vector component can assume a greater number of values (in this case 7, as opposed to 3). The number of values is constant for each person, which is not the case in the spell representation, but one "pays" for this by losing the date at which the change takes place. Obviously, one can remedy some of the disadvantages of the vector approach by choosing a shorter time period, for example, weekly rather than monthly. This increases the number of vector components.

Generally, the choice between representations depends on the unit of analysis. For example, if the person is the unit of analysis, the vector representation is probably more useful. One then has to decide on the time period represented by each component. There are trade-offs required. However, if one is looking, for example, at durations of unemployment, the unit of analysis is the unemployment spell, and the spell representation is probably preferable. One can always convert a spell representation into a vector representation. The other direction is also possible except that the start and end dates of spells are limited by the time frame used in the vector.

The SLID data model contains a mixture of spell and vector representations depending on what is perceived to be the major use of the variable. This does imply that data manipulation will be required for data users wishing a representation different from that provided in the data model.

Family / household data

When embarking on longitudinal analysis for the first time, a common thought is: "I would like to see what happens to a family over time.". This may seem reasonable and conceptually simple, but soon breaks down when one starts to describe the data required. What is a "longitudinal family" when its composition changes? The same problem exists for households over time. SLID will not be defining longitudinal families, but will be providing, for each person, family and household characteristics for that person. For example, one can study how a person's family income changes over time in relation to his/her labour market activity. The key difference is that the unit of analysis is the person and not the family. For more information on family and household data, see SLID Research Paper 94-06 *SLID Household and Family Variables.*

Geography of residence

Geography variables present similar types of issues as family variables, but the focus is slightly different, so they are discussed separately. As with other variables which can change over time, geography of residence must contain some representation of time.

However, it is slightly more complicated than the spell versus vector decision. Many analysts, such as those in provincial governments, are interested in subnational data. This seems like a straightforward request. However, one must first answer this question:

C Are you interested in a particular point in time or in a period of time?

If it is a point in time, then defining the population of interest is unambiguous. However, generally users are interested on at least a one year time frame. In this case, other questions must be answered:

C Are you interested in a particular point within the period (the end of the time frame, the beginning of the time frame, at another point in the time frame?), or in all persons living in the area of interest at any point in the time frame? (SLID will have information on moves each year.)

The answers to these questions may depend on the analysis. In any case, the results will differ (although perhaps not substantially) depending on how the population of interest is defined.

Following rules

Section 2 contained a very short description of the SLID following rules. However, it should not be surprising that the decisions regarding which persons get interviewed for each data collection period affect the survey data model.

Changes in key variables: date of birth / sex

Values for certain variables are essential to collection as they determine eligibility for certain questions. For most household surveys (SLID is no exception.), age (derived from date of birth) and sex are such variables. However, in a longitudinal survey, it is possible that errors can be corrected as part of a subsequent interview. This is an advantage in many ways, but also presents difficulties. Using SLID as an example, the annual labour interview is asked to everyone aged 16 or over at the end of the reference year. Assume that based on a person's date of birth, we determine that someone is 20 years old and therefore eligible for the interview. We collect the data. Then, at the next interview, we find out that the year of birth is incorrect, and the person was really only 12 at the time of the previous interview. Should we go back and delete the collected data or leave it? It is tempting to say "Yes", but suppose that the first year of birth was correct and it subsequently gets recorrected. We have then lost the data which we now want. Another possibility is to not allow any changes to the initially collected values. The rationale for this is that cross-sectional surveys have response errors to these variables, but they are not detected as the information is only collected once. However, in some cases it puts the interviewer in an awkward position of explaining to a respondent why we have not updated a person's date of birth or sex. Changes in these key variables are, hopefully, a rare occurrence and any decision will therefore have little impact on the use of the data. But a decision is required. SLID has decided to keep all collected data on the base, and to keep the values for date of birth and sex up to date; i.e., to believe all updates. A data user will then have to screen data based on the age and sex eligibilities.

Deferred May interview

While SLID contacts respondents twice annually, they can be considered to be two parts of one interview. They collect different information and both use the previous calendar year as the reference period. Income data are collected in May to improve the data quality, as this is the time when people have just submitted their income tax forms and are most knowledgeable about their sources of income. However, this May interview means that changes in household composition are collected twice a year.

In theory, SLID is to collect data from all longitudinal respondents and all persons living with longitudinal respondents at the end of December of the reference year. Collecting the labour information toward the end of January already means some changes can occur. But clearly, far more changes will take place between the end of the year and the middle of May when the income information is collected. The following possibilities must be planned for:

C Household composition changes between May interview and January interview - most of these will have occurred prior to December 31 but some will be after December 31

C Household composition changes between January interview and May interview

C In cases where no data collection takes place in May (unable to contact and nonresponse), household composition changes between one January interview and the next January interview; some will be before December 31 and some will be after

C In cases where no data collection takes place in January (unable to contact and nonresponse), household composition changes between one May interview and the next May interview; some will be before December 31 and some will be after

C Changes not being identified when they should be; for example, a person first reported in the May interview who moved in prior to December 31, but who was not reported in the January interview. This could result from confusion on the part of the respondent or interviewer, thinking that this person should not be identified. It could also result from a real ambiguity as to whether the person had actually moved in or not at the time of the January interview.

C Tracing of longitudinal persons may result in conflicting dates, making it impossible for SLID to determine the actual situation. For example, assume that a husband separates from his wife and moves in with his brother. The husband and wife are longitudinal respondents. In the interview following the separation, the wife reports the date when the husband moved out. He is traced, and the brother is asked when the

> husband moved in. Assume the wife reports a date which is a month earlier than the brother. Which is the true situation: the husband was in a third household for one month, recall error of the dates, the husband moved gradually without a clear moving date?

To the extent that the data reporting allows, as part of processing, SLID will attempt to establish the correct dates, ensuring that the existence of data for a reference year is consistent with the theoretical intent of the data collection.

Dependent interviewing - what to do with denials

Dependent interviewing is using data collected at a previous interview to help a respondent with the current interview. One example of SLID's use of dependent interviewing is the "feeding back" of the names of all employers for whom the respondent was working at the end of the previous reference period. For example, a person reports that at the end of Year 1, he was working for ABC Company and DEF Inc. As part of the labour interview for Year 2, the person is asked to confirm that he had these two employers at the beginning of Year 2. Without this feedback, it is possible that the person could forget one, particularly one that ended early in the year. The disadvantage of dependent interviewing is that the confirmation process could lead to a denial. What does one then do? To maintain a good relationship with the respondent, the interviewer must accept the information given; i.e., the denial. However, a decision remains for processing: Which of the two responses was correct? There does not appear to be a general rule to follow; the decision depends on the particular data item.

Processing considerations

This is primarily a common sense item, applicable to all surveys. One must have knowledge of how the data are to be processed when developing a data model. In that way, the data modellers can ensure that the data can be processed in the intended manner, and that the data model does not introduce restrictions. One simple example is the decision on whether to keep or overwrite the collected data when editing. If the raw data are simply a step in the process to producing an output file, then they can be overwritten (obviously, backup files would be kept). If a user wishes to do a comparison of a particular data variable before and after editing, then the data model should allow both versions of the variable to be stored on the same database.

8. SLID Data Model: Basic characteristics and future directions

A relational database has been adopted for storing SLID data. Several factors led to this decision:

C Several different units of analysis. The use of a relational database will reduce the file size and data redundancies.

C SLID data model should include all survey data, and not just those data required for analysis (files related to collection such as contact information on respondents are part of the SLID data model although they are obviously not available to data users)

C Not all respondents are eligible for all years. For joiners, data for years prior to joining the sample are not available. Persons under 15 have very little data collected, but some persons who are under 15 at the beginning of the six-year survey period will turn 15 in a subsequent survey year, and will become eligible for data collection.

The SLID data model currently has 44 entities. It is possible that this number will change slightly, either up or down, but will not change greatly until major survey content changes are introduced. One major content addition planned for collection in 1998 is information on a person’s assets and debts. This will undoubtedly require additional entities in the data model.

Some modification in variables will continue for some time, primarily due to the addition of derived variables. All suggested derived variables are added to the data model. Some of these will turn out to be impossible to calculate and will be dropped from the data model. Presently there are approximately 700 variables in the data model.

When using a relational database, the number of variables can be misleading. For example, one variable identifies whether in the context of a specific job, the person is a manager. This information is collected for every job for every year in which the person holds the job. Thus several values will be stored for every person, but the variable is only counted once.

The SLID data model contains four basic types of entities:

C Fixed entities are those whose values do not change. An example is the entity which contains information such as mother tongue, country of birth, and ethnic background.

C Annual entities are those which cover one calendar year or which contain point in time data which are collected only once a year. An example of the first type of annual entity is the one which stores the amount of income received from various sources for a given year. An example of the second type is the entity which contains information on persons reporting activity limitation and its effect on their labour market activity. The information relates to one point in time (theoretically, December 31 of the reference year). At the end of the six year survey period, six annual “snapshots” of these data will exist.

C Cumulative entities are annual entities whose value one year is derived by taking the value from the previous year and adding according to the data collected during the reference year. An example is the entity containing information on work experience. A person’s lifetime work experience at the end of Year 2 is the sum of the work experience at the end of Year 1 and the amount worked during Year 2.

C Spell entities are those for which one record corresponds to one spell. An example is the entity which stores information on a person’s marital status. It contains one record for every marital status for every person. One must combine all marital spells for a person to get information at the person level; i.e., a vector representation.

9. SLID microdata files and interface with users

Although the scope of this paper covers the development of the data model, a natural extension is to discuss the impact of this data model on data users. This section briefly discusses this topic.

As mentioned earlier, one concern when developing the SLID data model was the varied technical abilities of the data user community. Some users will be able to cope with a relational database environment, but the majority will not. In this sense “coping” means the varied learning curves for data users to manipulate the data, and not limitations of hardware and software. (These latter limitations will present some problems, but they are not felt to be great due to the rapidly changing technology.) It is important to provide additional tools to simplify use of the data:

Retrieval software

It is planned that SLID microdata files will be distributed on CD-ROM with “retrieval software”, which will provide an interface with the relational database. The user will pass through a series of screens which will allow the selection of the unit of analysis, the time period of interest, the population of interest, and the variables of interest. The retrieval software will then construct a flat file in ASCII format; in other words, a customized microdata file. In this way, the user need not learn how to manipulate data in relational databases. The created ASCII file can then be used in the user’s preferred analytical software.

Software for basic analysis

Another piece of software is also planned to accompany SLID microdata files. The exact logistics still remain to be determined. However, the concept is that this software would provide some basic analytical tools, such as summary statistics, cross tabulations, graphs and charts. Its purpose would be to provide a simple starting point for “getting a feel for” the data. This may even be sufficient for some users requiring only cross-tabulations.

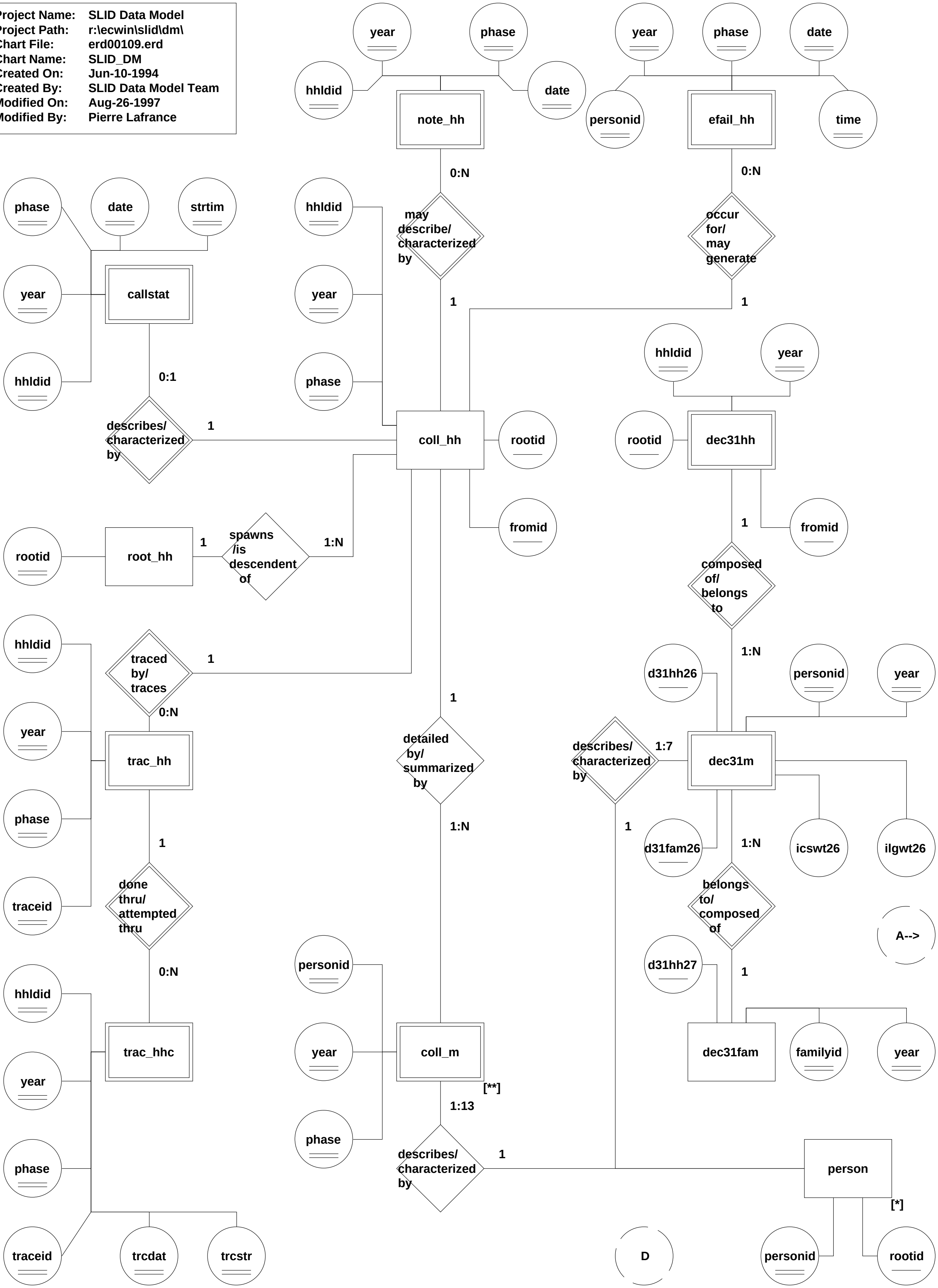
User workshops

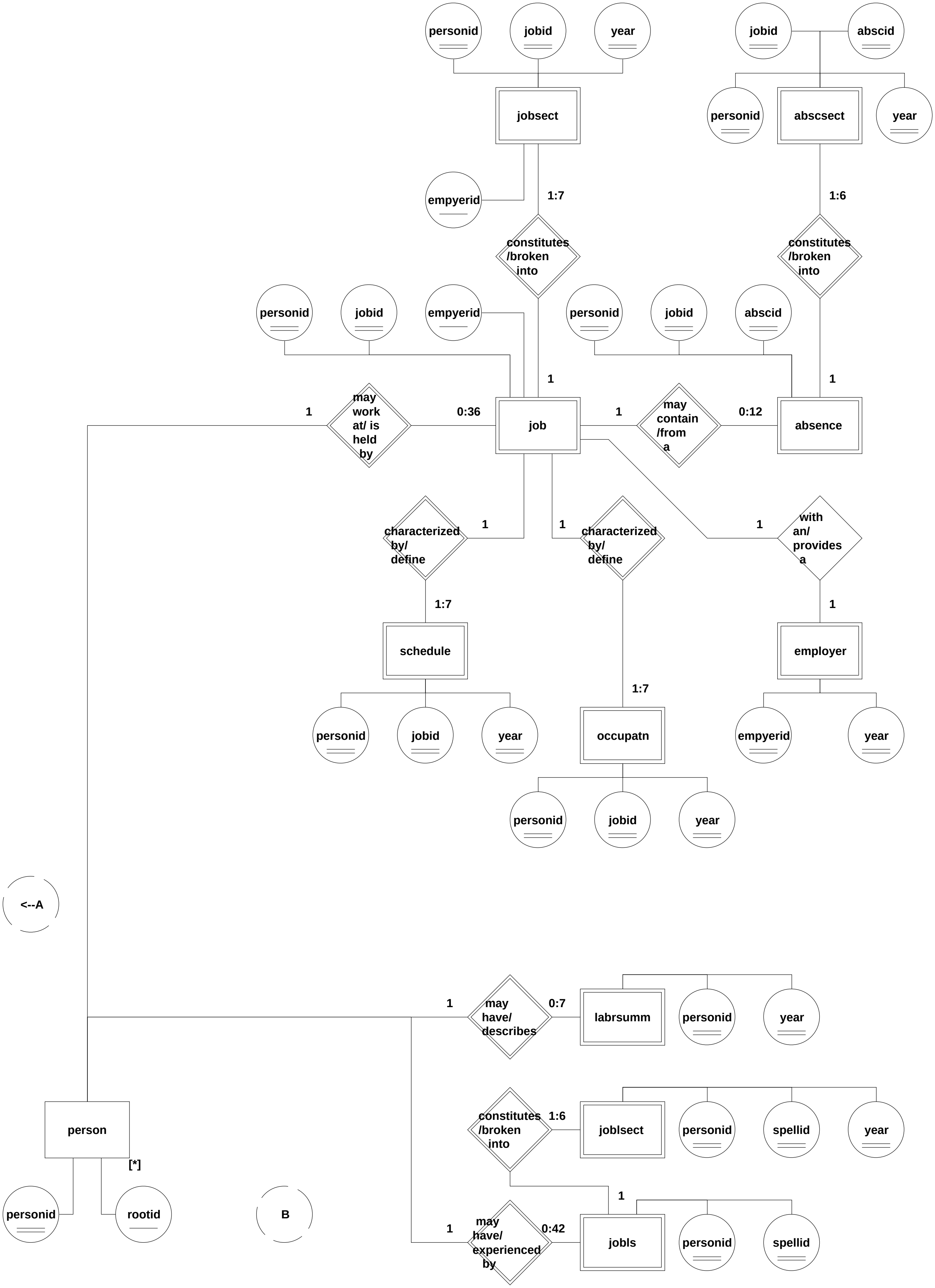
Often, questions arise only when one starts to use the data. Workshops which demonstrate some simple data retrieval and tabulations should provide a good basis for getting data users started.

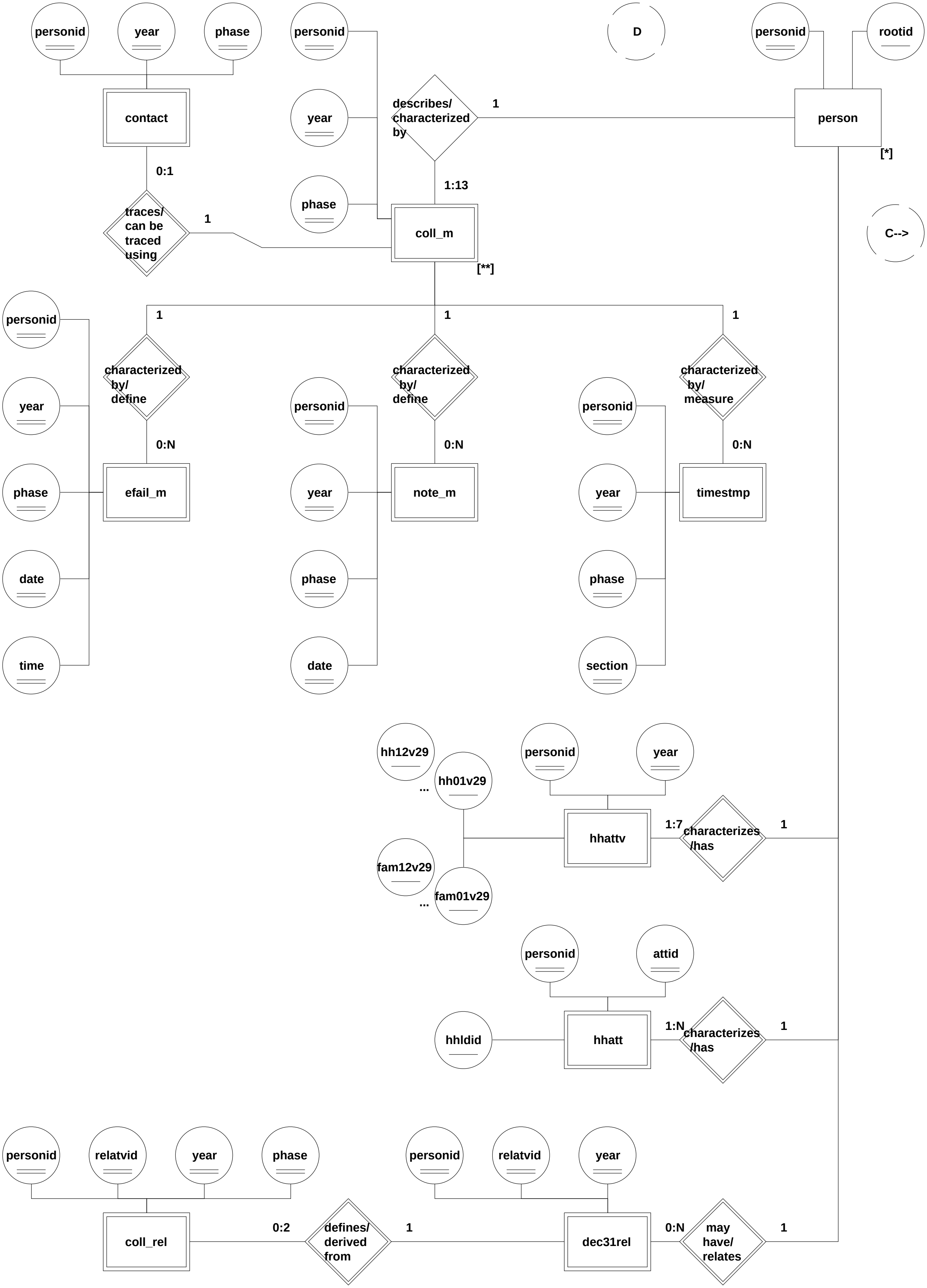
APPENDIX 1:

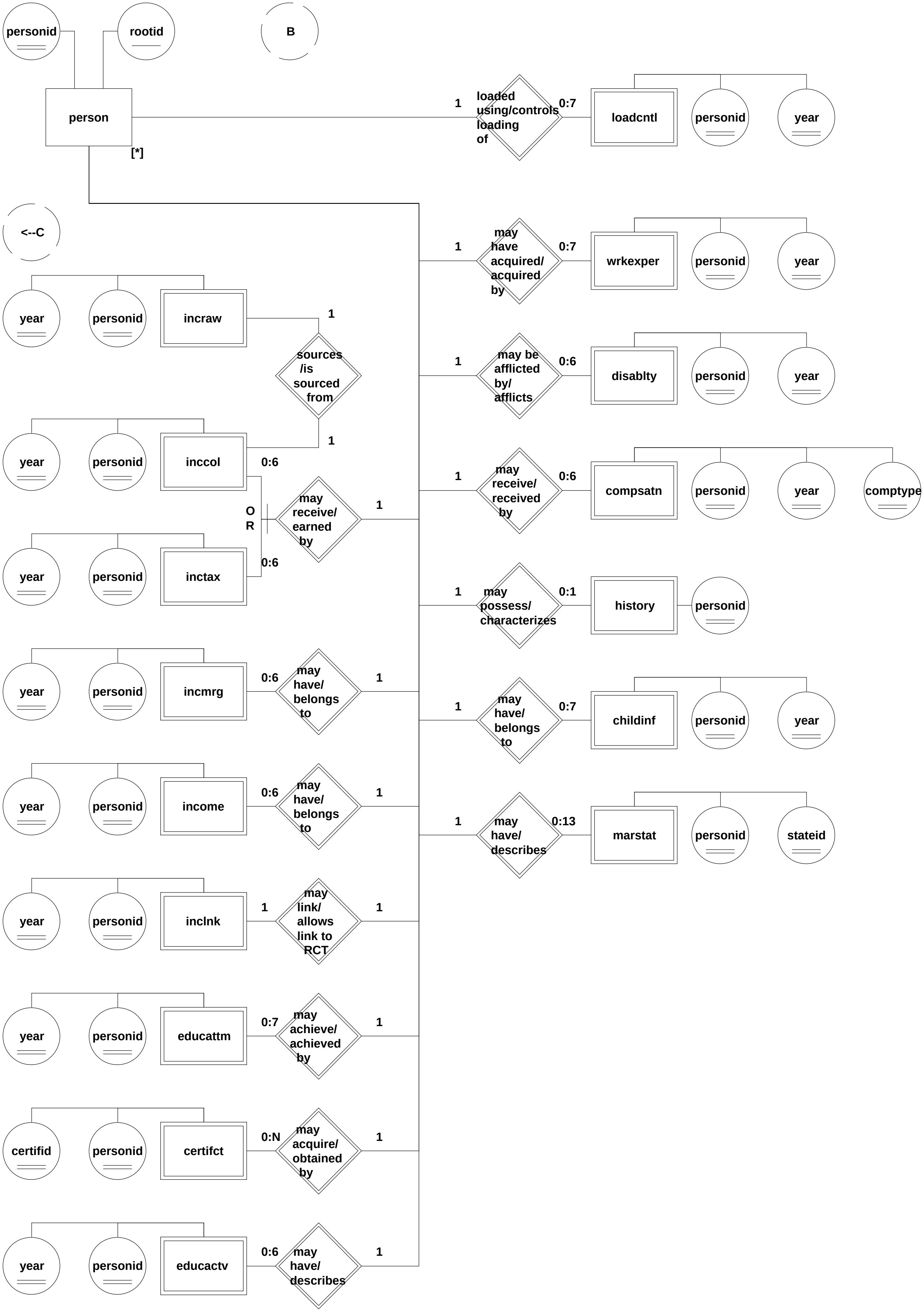
EXAMPLES OF ENTITY-RELATIONSHIP DIAGRAMS (ERDS) AND THE DATA DICTIONARY (DD) FROM THE SURVEY OF LABOUR AND INCOME DYNAMICS

Project Name:	**SLID Data Model**
Project Path:	**r:\ecwin\slid\dm\**
Chart File:	**erd00109.erd**
Chart Name:	**SLID_DM**
Created On:	**Jun-10-1994**
Created By:	**SLID Data Model Team**
Modified On:	**Aug-26-1997**
Modified By:	**Pierre Lafrance**




personid
jobid
year
jobid
abscid
jobsect
personid
abscsect
year
empyerid
1:7
1:6
constitutes /broken into
constitutes /broken into
personid
jobid
empyerid
personid
jobid
abscid
1
1
1
may work at/ is held by
0:36
job
1
may contain /from a
0:12
absence
characterized by/ define
1
1
characterized by/ define
1
with an/ provides a
1:7
1
schedule
employer
1:7
personid
jobid
year
occupatn
empyerid
year
personid
jobid
year
<--A
labrsumm
personid
year
1
may have/ describes
0:7
constitutes /broken into
1:6
joblsect
personid
spellid
year
person
[*]
1
personid
rootid
B
1
may have/ experienced by
0:42
jobls
personid
spellid


personid
year
phase
personid
D
personid
rootid
contact
year
describes/ characterized by
1
person
0:1
1:13
[*]
phase
traces/ can be traced using
1
coll_m
C-->
[**]
1
1
1
personid
characterized by/ define
characterized by/ define
characterized by/ measure
year
personid
personid
0:N
0:N
0:N
phase
efail_m
year
note_m
year
timestmp
date
phase
phase
time
date
section
hh12v29
personid
year
hh01v29
...
hhattv
1:7
characterizes /has
1
fam12v29
fam01v29
...
personid
attid
hhldid
hhatt
1:N
characterizes /has
1
personid
relatvid
year
phase
personid
relatvid
year
coll_rel
0:2
defines/ derived from
1
dec31rel
0:N
may have/ relates
1


personid
rootid
B
person
[*]
1
loaded using/controls loading of
0:7
loadcntl
personid
year
<--C
1
may have acquired/ acquired by
0:7
wrkexper
personid
year
year
personid
incraw
1
sources /is sourced from
1
1
may be afflicted by/ afflicts
0:6
disablty
personid
year
year
personid
inccol
0:6
1
may receive/ received by
0:6
compsatn
personid
year
comptype
O
R
may receive/ earned by
1
year
personid
inctax
0:6
1
may possess/ characterizes
0:1
history
personid
year
personid
incmrg
0:6
may have/ belongs to
1
1
may have/ belongs to
0:7
childinf
personid
year
year
personid
income
0:6
may have/ belongs to
1
1
may have/ describes
0:13
marstat
personid
stateid
year
personid
inclnk
1
may link/ allows link to RCT
1
year
personid
educattm
0:7
may achieve/ achieved by
1
certifid
personid
certifct
0:N
may acquire/ obtained by
1
year
personid
educactv
0:6
may have/ describes
1

SLID DATA MODEL:
ENTITY STRUCTURE

Entity: **JOB**

Alias: Entity 9

Purpose: Spell entity used to retain information about the characteristics of an employment spell or job associated with one employer. The employment spell may span more than one year.

Unit of Analysis: Person-Job

Key Fields:

PERSONID Character (12) Format: PPWFMMMMMMMM

Long name: Person ID

Description: Unique identifier for a person.

(Note. Suppressed on the public use file)

The format of this field is: PPWFMMMMMMMM

P - panel identifier
W - wave in which person entered sample
F - phase in which person entered sample
M - case identifier for person

Population: All persons

Processing: load from 12 byte expanded DMPRSNID from methodology person file.

Reserved Codes:

999999999996	Not in Sample
999999999997	Don’t Know
999999999998	Refusal
999999999999	Not Applicable

JOBID Numeric (2)

Long name: Job spell ID

Description: Unique identifier for a job or employment spell with an employer. The same employment spell that crosses over two reference years should retain the same "jobid". Two distinct employment spells with the same employer within the same reference year would each have a distinct "jobid".

Population: Persons aged 16-69 AND employed during the survey reference period

Range: 00:95

Reserved Codes:

96	Not in Sample
97	Don’t Know
98	Refusal
99	Not Applicable

Attributes:

EMPYER9 Character (14) Format: PPWFMMMMMMMMJJ

Long name: Employer ID

Description: Unique identifier for an employer. All distinct jobs for a person with the same employer will be assigned the same employer identifier which will correspond to the employer identifier associated with the first job.

(Note. Suppressed on the public use file)

The format of this field is: PPPPPPPPPPPPJJ

P - person identifier
J - job identifier of first job held by a person with the employer

Population: Persons aged 16-69 AND had a job during the survey reference period

Reserved Codes:

99999999999996	Not in Sample
99999999999997	Don’t Know
99999999999998	Refusal
99999999999999	Not Applicable

STRDAT9 Date (8) Format: YYYYMMDD

Long name: Start date of job

Description: Start date of job.

All SLID "date" type variables can be viewed as a concatenation of the three sub-component fields year, month and day. The standard SLID reserved values are assigned, when necessary, to the component fields, and not to the complete date variable. So it is possible to have a mixture of valid values and reserved codes in a "date" type variable.

(Note. Day suppressed on the public use file)

Population: Persons aged 16-69 AND had a job during the survey reference period

Source: [DATES-Q2T1]
[DATES-Q2T2]
[DATES-Q3], [DATES-Q5], [DATES-Q6]

Range: 18000101:99951231

Reserved Codes:

96	Not in Sample
97	Don’t Know
98	Refusal
99	Not Applicable
9996	Not in Sample
9997	Don’t Know
9998	Refusal
9999	Not Applicable

STRDA_9 Character (3) Format: 3 byte array of format YMD

Long name: Date imputation flags

Description: Date imputation flags.

Population: Persons aged 16-69 AND had a job during the survey reference period

Codes:	0	collected date component
	1	imputed date component

ENDDAT9 Date (8) Format: YYYYMMDD

Long name: End date of job

Description: End date of job.

All SLID "date" type variables can be viewed as a concatenation of the three sub-component fields year, month and day. The standard SLID reserved values are assigned, when necessary, to the component fields, and not to the complete date variable. So it is possible to have a mixture of valid values and reserved codes in a "date" type variable.

(Note. Day suppressed on the public use file)

Population: Persons aged 16-69 AND had a job during the survey reference period

Source: [DATES-Q3], [DATES-Q8]
[DATES-Q10], [DATES-Q11], [DATES-Q12]

Range: 18000101:99951231

Reserved Codes:	96	Not in Sample
	97	Don’t Know
	98	Refusal
	99	Not Applicable
	9996	Not in Sample
	9997	Don’t Know
	9998	Refusal
	9999	Not Applicable

ENDDA_9 Character (3) Format: 3 byte array of format YMD

Long name: Date imputation flags

Description: Date imputation flags.

Population: Persons aged 16-69 AND had a job during the survey reference period

Codes:	0	collected date component
	1	imputed date component

FSTDAT9 Date (8) Format: YYYYMMDD

Long name: Start date with employer

Description: Date first worked for this employer.

All SLID "date" type variables can be viewed as a concatenation of the three sub-component fields year, month and day. The standard SLID reserved values are assigned, when necessary, to the component fields, and not to the complete date variable. So it is possible to have a mixture of valid values and reserved codes in a "date" type variable.

(Note. Day suppressed on the public use file)

Population: Persons aged 16-69 AND had a job during the survey reference period

Source: PRELIM-Q21A,Q21B
PRELIM-Q25A,Q25B
[DATES-Q7], [DATES-Q7A]

Range: 18000101:99951231

Reserved Codes:	96	Not in Sample
	97	Don’t Know
	98	Refusal
	99	Not Applicable
	9996	Not in Sample
	9997	Don’t Know
	9998	Refusal
	9999	Not Applicable

FSTDA_9 Character (3) Format: 3 byte array of format YMD

Long name: Date imputation flags

Description: Date imputation flags.

Population: Persons aged 16-69 AND had a job during the survey reference period

Codes:	0	collected date component
	1	imputed date component

ENDED9 Character (1)

Long name: Job ended

Description: Flag to indicate if job had ended by the end of the most current survey reference period of the data file.

Population: Persons aged 16-69 AND had a job during the survey reference period

Source: [DATES-Q3], [DATES-Q8], [DATES-Q10]
[DATES-Q11], [DATES-Q12]
JOB.ENDDAT9

Codes:	1	Yes
	2	No
Reserved Codes:	6	Not in Sample
	7	Don’t Know
	8	Refusal
	9	Not Applicable

ENDTYP9 Character (1)

Long name: Type of job end

Description: Reason why job was ended in processing. This is an internal variable to be used for processing only.

Population: Persons aged 16-69 AND had a job during the survey reference period

Codes:	1	Job ended normally
	2	Job ended because the job was denied by the respondent
	3	Job ended because did not receive information about the job in subsequent collections - either because no feedback, non-response or because respondent is no longer eligible for the labour interview
Reserved Codes:	6	Not in Sample

Reserved Codes:	7	Don’t Know
	8	Refusal
	9	Not Applicable

REAEND9 Character (2)

Long name: Reason for job separation

Description: Reason why work came to an end.

Population: Persons aged 16-69 AND had a job separation during the survey reference period AND job has ended

Source: [DATES-Q13], [DATES-Q13A]
[DATES-Q13A1], [DATES-Q13A2], [DATES-Q13B]

Codes:	01	Own illness or disability - work related
	02	Own illness or disability - not work related
	03	Caring for own children
	04	Caring for elder relative(s)
	05	Other personal or family responsibilities
	06	School
	07	Found new job
	08	Move to a new residence
	09	Poor pay
	10	Not enough hours of work
	11	Too many hours of work
	12	Poor physical conditions (bad ventilation, too noisy, etc.)
	13	Sexual harassement
	14	Personnel conflict with employer/other employees
	15	Work too stressful
	16	Company moved
	17	Company went out of business
	18	Seasonal nature of work
	19	Layoff/Business slowdown (not caused by seasonal conditions)
	20	Labour dispute
	21	Dismissal by employer
	22	Temporary job/contract ended
	23	To concentrate on other job
	24	Retirement
	25	Other
Reserved Codes:	96	Not in Sample
	97	Don’t Know
	98	Refusal
	99	Not Applicable

TYPJS9 Character (1)

Long name: Type of job separation

Description: Type of job separation.

Population: Persons aged 16-69 AND had a job separation during the survey reference period AND job has ended

Source: JOB.REAEND9

Codes:	1	Voluntary
	2	Involuntary
Reserved Codes:	6	Not in Sample
	7	Don’t Know
	8	Refusal
	9	Not Applicable

CLWKR9 Character (2)

Long name: Class of worker

Description: Most current class of worker for this job.

Population: Persons aged 16-69 AND had a job during the survey reference period

Source: PRELIM-F05Q76
PRELIM-Q29,COW2_COD
[CHAR-Q3], [CHAR-Q3A], [CHAR-Q3B], [CHAR-Q3C]
JOBSECT.CLWKR1

Codes:	01	Paid worker
	02	Unpaid family worker
	03	Self-employed with paid help (Incorporated business)
	04	Self-employed with no paid help (Incorporated business)
	05	Self-employed with paid help (Not incorporated business)
	06	Self-employed with no paid help (Not incorporated business)
Reserved Codes:	96	Not in Sample
	97	Don’t Know
	98	Refusal
	99	Not Applicable

RCLWKR9 Character (2)

Long name: Recoded class of worker

Description: Most current recoded class of worker for this job. Owners of incorporated businesses have been included in the "Paid Worker" category. Also, a separate category has been created for government employees.

Population: Persons aged 16-69 AND had a job during the survey reference period

Source: PRELIM-F05Q76 (job 1)
PRELIM-Q29, COW2_COD (job 2)
[CHAR-Q3], [CHAR-Q3A], [CHAR-Q3B], [CHAR-Q3C]
JOBSECT.RCLWKR1

Codes:	01	Paid worker
	02	Paid worker, government
	04	Self-employed with paid help (unincorporated business)
	05	Self-employed without paid help (unincorporated business)
	06	Unpaid family worker
Reserved Codes:	96	Not in Sample
	97	Don’t Know
	98	Refusal
	99	Not Applicable

HOWOBT9 Character (2)

Long name: How job was obtained

Description: How job was obtained.

Population: Persons aged 16-69 AND had a job during the survey reference period AND job started after Jan 1/93 AND paid worker

Source: [CHAR-Q4]

Codes:	01	Contacted employer directly
	02	Friend or relative
	03	Placed or answered newspaper ad
	04	Employment agency
	05	Referral from another employer
	06	Contacted directly by employer
	07	Union
	08	Other
Reserved Codes:	96	Not in Sample
	97	Don’t Know
	98	Refusal
	99	Not Applicable

OFRDAT9 Date (8) Format: YYYYMMDD

Long name: Date job offer received

Description: Date the job was offered.

All SLID "date" type variables can be viewed as a concatenation of the three sub-component fields year, month and day. The standard SLID reserved values are assigned, when necessary, to the component fields, and not to the complete date variable. So it is possible to have a mixture of valid values and reserved codes in a "date" type variable.

(Note. Day suppressed on the public use file)

Population: Persons aged 16-69 AND had a job during the survey reference period AND job started after Jan 1/93 AND paid worker

Source: [CHAR-Q5]

Range: 18000101:99951231

Reserved Codes:	96	Not in Sample
	97	Don’t Know
	98	Refusal
	99	Not Applicable
	9996	Not in Sample
	9997	Don’t Know
	9998	Refusal
	9999	Not Applicable

OFRDA_9 Character (3) Format: 3 byte array of format YMD

Long name: Date imputation flags

Description: Date imputation flags.

Population: Persons aged 16-69 AND had a job during the survey reference period

Codes:	0	collected date component
	1	imputed date component

NBOCCU9 Numeric (2)

Long name: Duties changed-job

Description: Number of kind of work, most important activities and duties descriptions for this job.

(RESERVED FOR FUTURE USE)

Population: Persons aged 16-69 AND had a job during the survey reference period

Source: [CHAR-Q11]

Range: 01:95

Reserved Codes:	96	Not in Sample
	97	Don’t Know
	98	Refusal
	99	Not Applicable

OCCHG9 Character (1)

Long name: Duties changed-job

Description: Flag to indicate whether respondent reported that kind of worked changed since job started.

Population: Persons aged 16-69 AND had a job during the survey reference period

Codes:	1	Yes
	2	No
Reserved Codes:	6	Not in Sample
	7	Don’t Know
	8	Refusal
	9	Not Applicable

NBSCHD9 Numeric (2)

Long name: No. work schedules-job

Description: Number of work schedules for this job.

Population: Persons aged 16-69 AND had a job during the survey reference period

Source: [CHAR-Q27], [CHAR-Q32]

Range: 00:95

Reserved Codes:	96	Not in Sample
	97	Don’t Know
	98	Refusal
	99	Not Applicable

NBABS9 Numeric (2)

Long name: No. absences-job

Description: Excluding paid vacation, number of times respondent was absent from this job for more than one week.

Population: Persons aged 16-69 AND had a job during the survey reference period

Source: [CHAR-Q46] (for t1 or t3)
[CHAR-Q48], [CHAR-Q48A], [CHAR-Q48B]
[CHAR-Q49], [CHAR-Q49A]

Range: 00:95

Reserved Codes:	96	Not in Sample
	97	Don’t Know
	98	Refusal
	99	Not Applicable

NBMTWK9 Numeric (2)

Long name: No. mos worked-job

Description: Total months in which some work was done at this job excluding months where person was absent from the job all month..

Population: Persons aged 16-69 AND had a job during the survey reference period

Source: [DATES-Q2T1], [DATES-Q2T2], [DATES-Q3]
[DATES-Q5], [DATES-Q6], [DATES-Q10], [DATES-Q12]
JOB.STRDAT9
JOB.ENDDAT9
JOB.ENDED9

Range: 00:95

Reserved Codes:	96	Not in Sample
	97	Don’t Know
	98	Refusal
	99	Not Applicable

JOBDUR9 Numeric (3)

Long name: Duration of job in mos

Description: Duration of job (expressed in months).

Population: Persons aged 16-69 AND had a job during the survey reference period

Source: JOB.STRDAT9
JOB.ENDDAT9

Range: 001:995

Reserved Codes:	996	Not in Sample
	997	Don’t Know
	998	Refusal

CONATT9 Character (1)

Long name: Continuous employer

Description: Flag to indicate whether attachment to this employer has been continuous since first started working for this employer.

Population: Persons aged 16-69 AND had a job during the survey reference period AND job continuous since start date

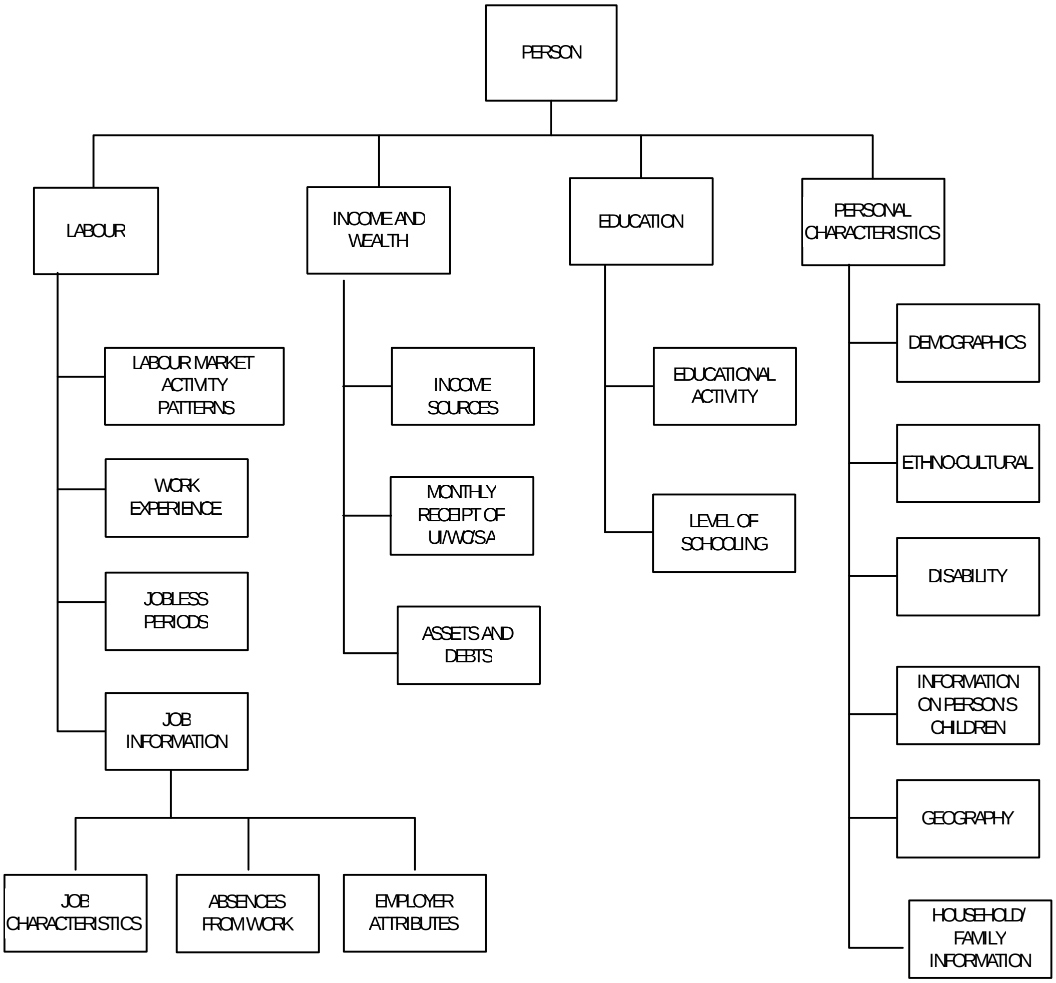
Source: JOB.STRDAT9
JOB.FSTDAT9

Codes:	1	Yes
	2	No
Reserved Codes:	6	Not in Sample

Reserved Codes:	7	Don’t Know
	8	Refusal
	9	Not Applicable

APPENDIX 2: 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 equivalents

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)

- start and end dates, first date ever worked for this employer
- wage
- work schedule (hours and type)
- benefits
- union membership
- occupation
- supervisory and managerial responsibilities

- class of worker
- tenure
- 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
- public or private sector

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
- Workers' Compensation

Assets and debts

Information may 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
- duration of current marital status
- 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 census metropolitan area 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/family (e.g., age, sex, relationship, income, annual hours worked)
- relevant low-income cutoff
- family events (separation, death, birth)
- dwelling type and tenure