WP 18

ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Madrid, Spain, 20-22 October 2003)

Topic (ii): Developments related to new methods and techniques

CURRENT AND FUTURE APPLICATIONS OF CANCEIS AT STATISTICS CANADA

Supporting Paper

Submitted by Statistics Canada[1]

I. INTRODUCTION

1. Many minimum change imputation systems are based on the approach proposed by Fellegi and Holt (1976). For example, CANEDIT and GEIS/Banff (Statistics Canada, 2003) at Statistics Canada, and DISCRETE and SPEER at United States Bureau of the Census all use, or had as their starting point, the Fellegi/Holt imputation methodology. In the 1996 Canadian Census of Population, a somewhat different approach was used successfully to impute for non-response and inconsistencies for the demographic variables of all persons in a household simultaneously. The method used is called the Nearest-neighbour Imputation Methodology (NIM). This implementation of the NIM allowed, for the first time, the simultaneous hot deck imputation of qualitative and quantitative variables for large E&I problems. In Bankier (1999), an overview of the NIM algorithm is provided.

2. The main difference between the NIM and the Fellegi/Holt imputation methodology is that the NIM first finds donors and then determines the minimum number of variables to impute based on these donors. The Fellegi/Holt methodology determines the minimum number of variables to impute first, and then attempts to find donors. Reversing the order of these operations confers significant computational advantages to implementations of the NIM while still meeting the well-accepted Fellegi/Holt objectives of minimum change and preserving sub-population distributions. The NIM, however, in its present form, can only be used to carry out imputation using donors while the Fellegi/Holt can be used with any imputation methodology.

3. For the 2001 Census, a more generic implementation of the NIM was developed. It is called the CANadian Census Edit and Imputation System (CANCEIS). Besides the demographic variables, it was used in 2001 to perform E&I for the labour, mobility, place of work and mode of transport variables. This corresponds to 40% of all variables on the 2001 Census questionnaire. The SPIDER E&I system (System for Processing Instructions from Directly Entered Requirements, which has been used since 1981) processed the other 60% of the Census variables. For the 2006 Canadian Census, CANCEIS will process all census variables. CANCEIS has also been used by the Canadian Census of Agriculture Coverage Evaluation Survey and is used by the annual Canadian Survey of Household Spending.

[1] Prepared by Michael Bankier, Social Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6. Mike.Bankier@statcan.ca.

4. Prior to the development of CANCEIS, enhancements to the NIM were implemented in prototype software. This software was used to process variables in the 2000 Brazilian Census and the 2001 Swiss Census. In addition, the 2001 Italian Census, having studied CANCEIS, will use a similar approach in their imputation methodology. CANCEIS has also been recently provided to the national statistical offices of the United Kingdom, Brazil and Peru for their evaluation.

5. In the 2001 Census, besides nearest neighbour imputation, SPIDER also performed deterministic imputation and the derivation of new variables. In 2001, CANCEIS performed nearest neighbour imputation but not the other two functions. For 2006, CANCEIS will be extended to derive variables and perform deterministic imputation.

6. In the 2001 Census, the SPIDER modules were implemented using approximately five thousand machine-readable decision logic tables (DLTs). Because of budget and time constraints, the conversion from SPIDER to CANCEIS must be done as efficiently as possible. It is planned to initially convert SPIDER modules to CANCEIS with few changes. After all modules are converted, some enhancements may be done. A few modules, however, will have to be rewritten because of major changes to Census questions (e.g. education).

7. Section II describes the new CANCEIS Windows interfaces for specifying input data files and DLTs and for submitting CANCEIS jobs. Section III discusses other interfaces which will be used to translate SPIDER DLTs into CANCEIS DLTs. Finally, Section IV briefly describes other improvements being made to CANCEIS and provides some concluding remarks.

II. CANCEIS WINDOWS INTERFACES FOR SPECIFYING INPUT FILES AND DLTS

8. CANCEIS processed 40% of the 2001 Census variables on personal computers (PCs) using Windows NT. Processing these on the mainframe computer would have cost almost four hundred thousand dollars Canadian. CANCEIS ran under DOS with no Windows interface in 2001. Most CANCEIS input files were generated using a text editor. The DLTs were created using an Excel spreadsheet with macros for importing and exporting the DLTs as text files. CANCEIS jobs were submitted using DOS batch files.

9. For 2006, Windows interfaces for CANCEIS have been created to simplify its use. It should be noted that the use of these interfaces is optional. The primary interface (called the CANCEIS Interface) helps the user to submit jobs plus specify most of the input files required by CANCEIS. The second interface (called the DLT Editor) makes it easier to specify the DLTs. These two interfaces will be illustrated using a simple DLT in Table 1 below. This DLT has one edit rule involving the single coded variable MARST (marital status) and the single discrete variable AGE. It indicates that a person who was ever married and under the age of 15 should fail the edit. This edit is applied to each person in a household.

Table 1: A Simple Example of a CANCEIS DLT

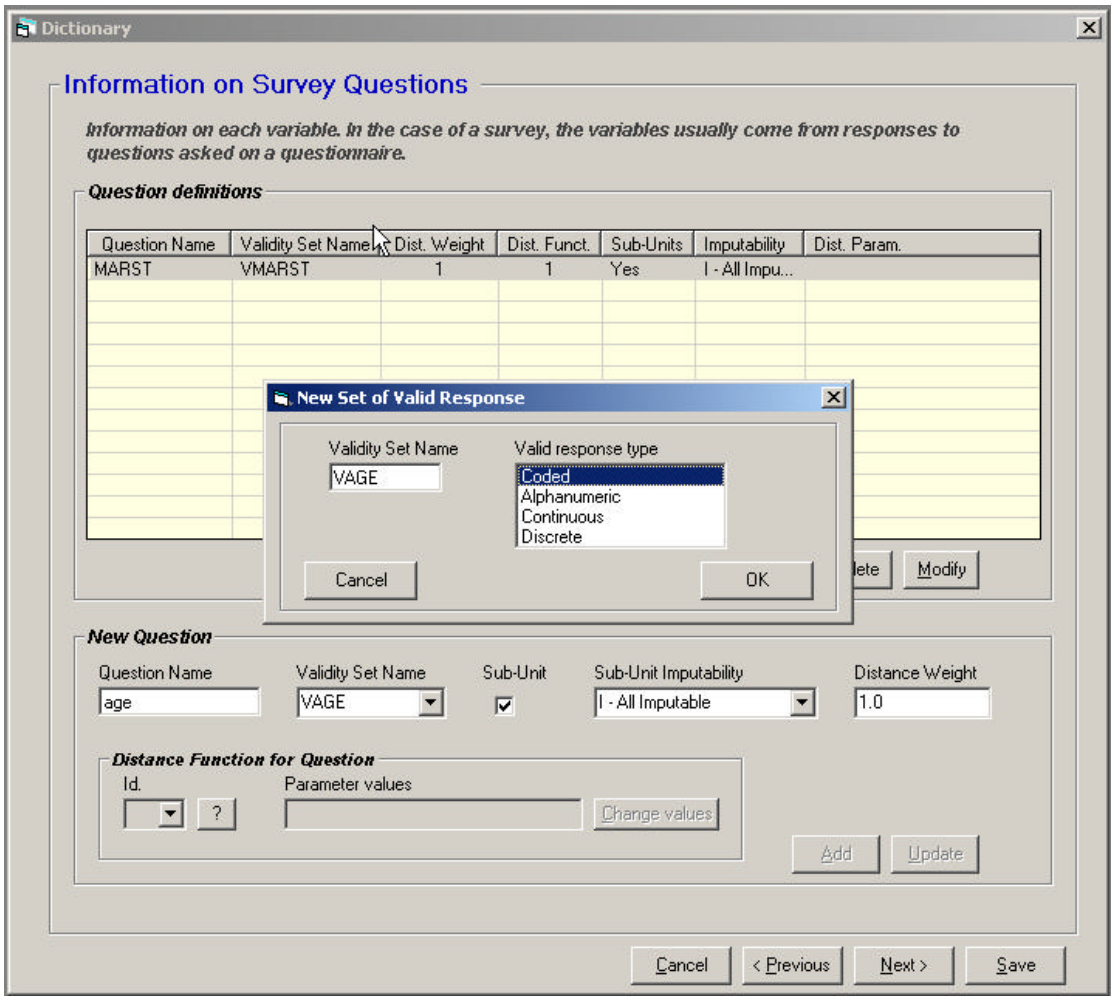
```
* *************************
* Description
*
*
* Date
* Author
* *************************
% DLT Name:mbtest

% Strata:                    2
% Purpose:                   Consistency
% Type:                      Conflict
% Symmetry:                  YES
% Sub-unit Start position:   1
% Sub-unit End position:     2

@ MARST(#1) = CLASS(evermarried)      ;Y;
@ AGE(#1) < 15                        ;Y;
```

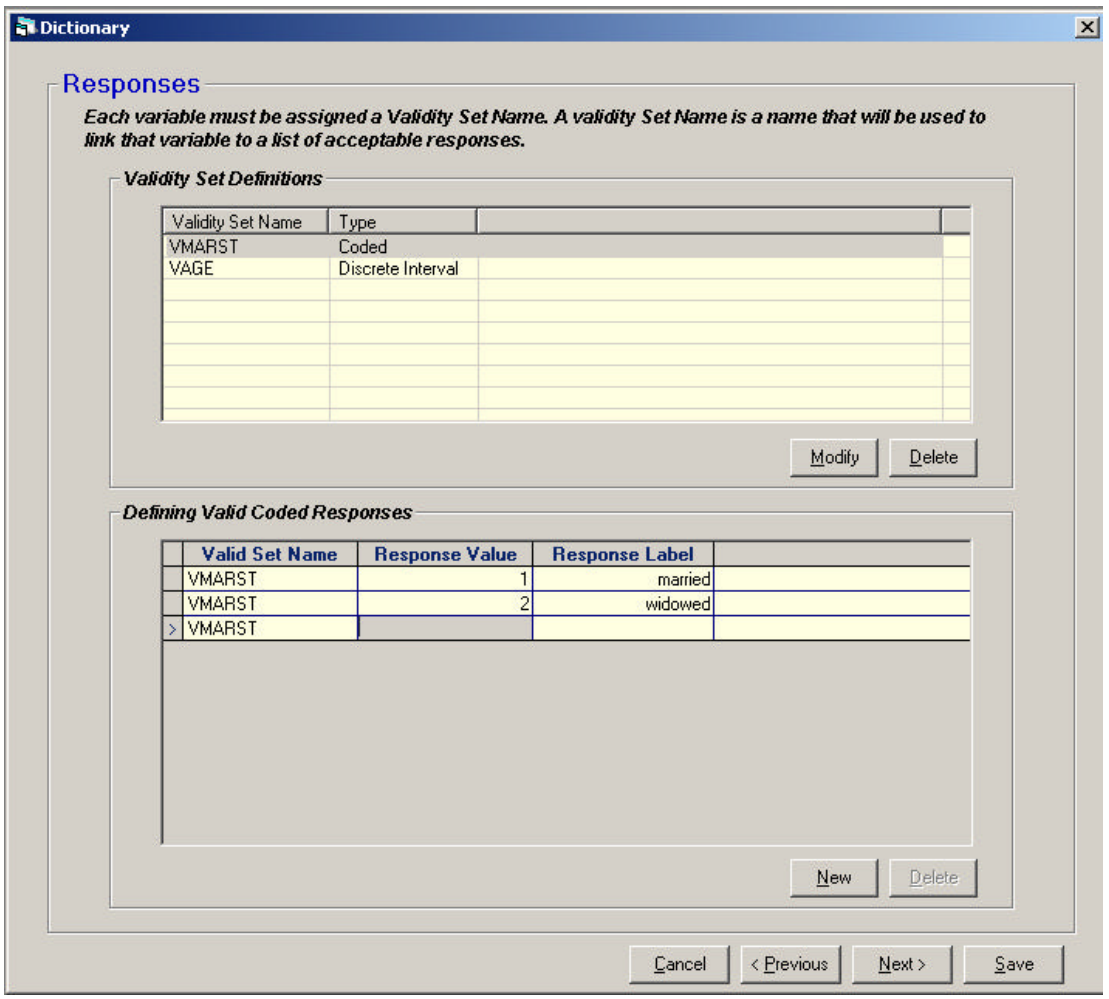
10. Before CANCEIS can use this DLT, the variables, their possible responses and classes of responses must be defined in a data dictionary. The Data Dictionary Wizard in the CANCEIS Interface requires the user to fill out a number of forms to do this. Figure 1 shows the form where the user lists all the questions in the survey and defines the name of a validity set (a list of valid responses) associated with each variable. Figure 1 shows that a variable MARST has already been defined and that it has the validity set VMARST associated with it. A second variable AGE is being defined and it has the validity set VAGE associated with it. VMARST was defined as a Coded type validity set while VAGE will be defined as a Discrete type validity set. By indicating "Yes" for sub-units for MARST and AGE, this indicates that these variables exist for each sub-unit (each person within a household) of the unit (the household).

Figure 1: Listing of Questions (Variables) and Initial Definitions of a Validity Sets



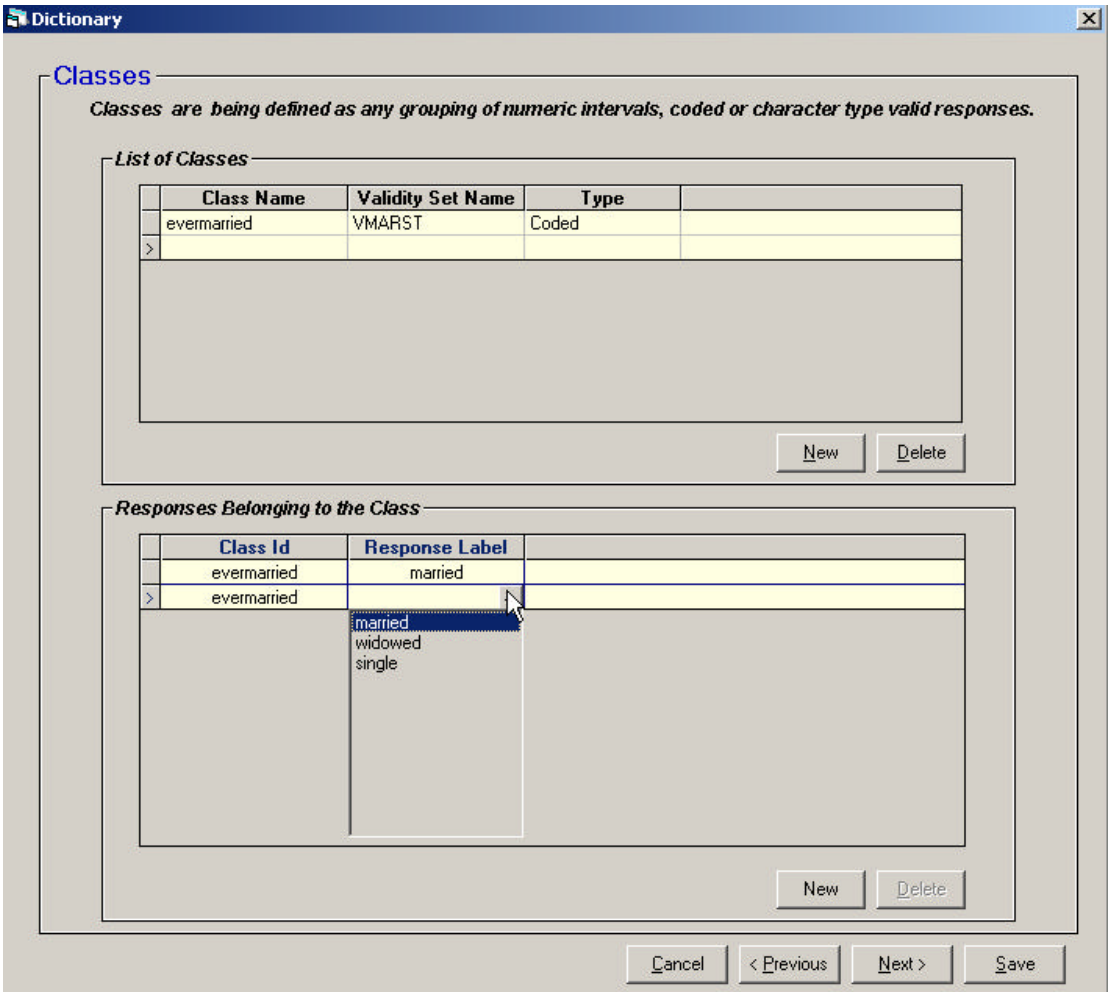
11. Figure 2 displays the form where the user specifies the valid values for each validity set. It can be seen that two of the valid values for VMARST are 1 and 2 (the responses for coded variables are assumed to be integers) and they have the labels “married” and “widowed” associated with them. The valid values for VAGE will be defined in terms of an interval, e.g. integers in the range 0 to 121 are considered valid.

Figure 2: Defining Valid Responses for a Validity Set



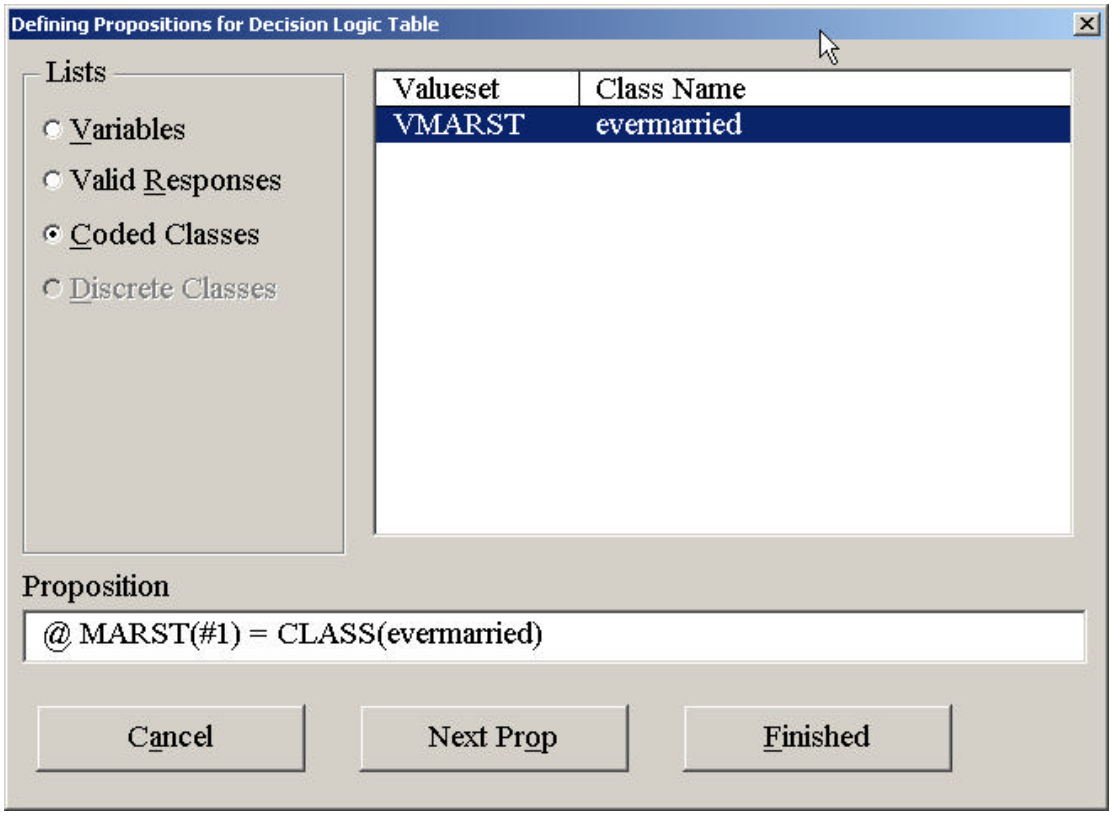
12. Figure 3 shows the form where the user specifies classes of valid responses for a validity set, in this case “evermarried”. “evermarried” will equal the set of responses “married” and “widowed” but not “single”.

Figure 3: Defining a Class of Responses for a Validity Set



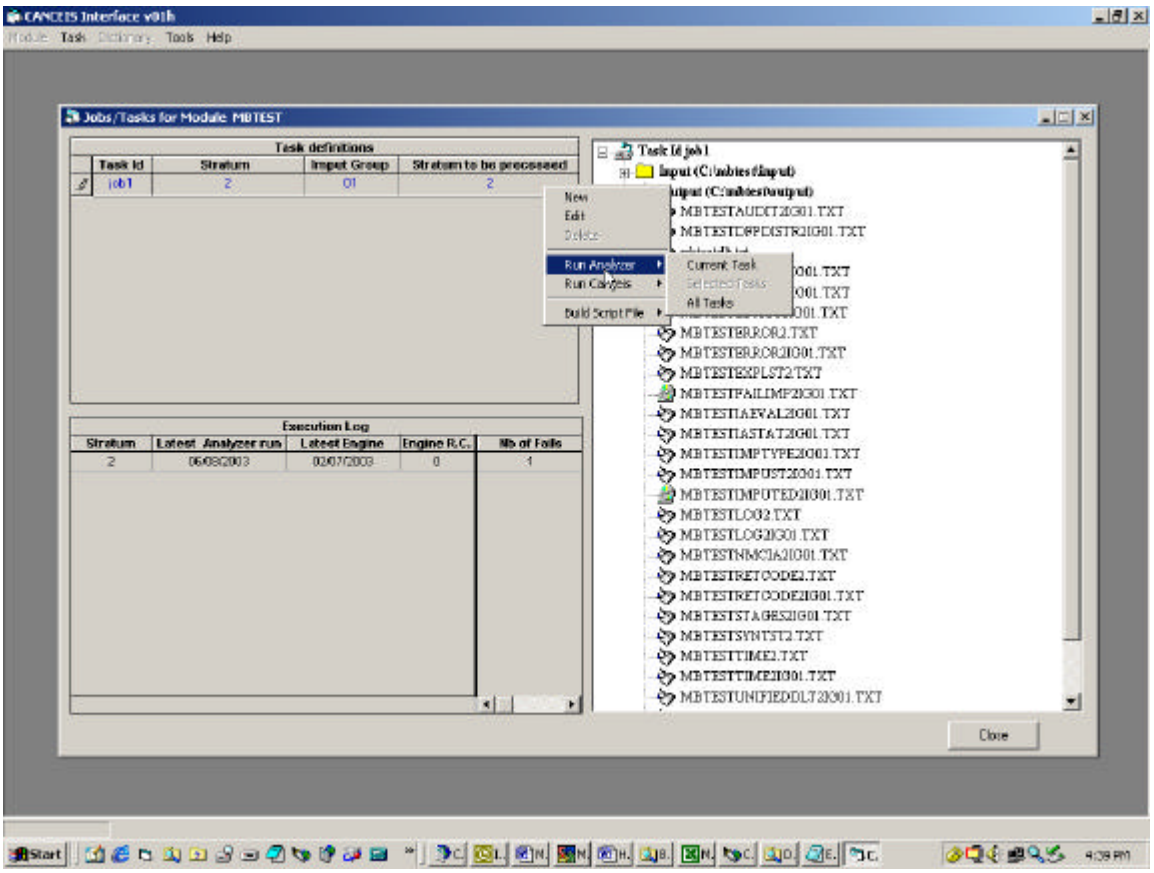
13. Having defined the variables, validity sets and classes of valid responses in a data dictionary, the variables, labels and classes can be used when specifying edits in DLTs. Figure 4 shows a form used in the DLT Editor to help define the proposition `MARST(#1) = CLASS(evermarried)`. The user has access to the lists of variables, valid responses and classes and can insert them into the proposition by double-clicking on one. This avoids typing in the variable names, responses and classes and thus eliminates typing errors. The DLT Editor also contains a Header editor (not illustrated) which makes it easier to fill in parameters (such as `% Symmetry:`) which precede the edits.

Figure 4: Proposition Editor Within DLT Editor



14. The CANCEIS Interface, besides helping the user to define data dictionaries, also makes it easier to submit and monitor jobs. Figure 5 shows the form used to do this. The Analyser can be run to check the DLTs for syntax errors followed by the CANCEIS Engine to perform E&I. Also listed are previous jobs run plus the input and output files from the current job. Script files can also be written which will allow a whole series of jobs to be run in sequence.

Figure 5: CANCEIS Interface for Submitting Jobs



III. WINDOW INTERFACES FOR TRANSLATING SPIDER DLTS TO CANCEIS DLTs

15. The CANCEIS DLT syntax, after it has been extended to allow deterministic imputation and the derivation of new variables, will be similar but not identical to that of SPIDER. To aid in the translation of DLTs from SPIDER to CANCEIS, a Windows interface (called the DLT Translator) has been created to automate certain aspects of the conversion. Table 2 shows a SPIDER DLT before translation while Table 3 shows the same DLT after the automated translation. Numbers at the end of lines have been eliminated. Some CANCEIS parameters have been added. Small changes to the syntax have been made. Additional manual clean-up of the DLT in Table 3 will be done within the DLT editor.

Table 2: SPIDER DLT Before Translation to CANCEIS

```
H NAME(ETHCKR) TYPE(A)                                                    0000000
*                                                                         0000000
*************************************************************************0000000
* This table originates from ASRCKR.  It is the fourth table of the    *0242046
* of the IBFN sub-module.  The tab checks if respondents provide any   *0000000
* Aboriginal ethnic origin in Q17.  Derived variable, ETHIND, is used  *8170000
* to identify Aboriginal ethnic origin.                                *5000000
* ETHIND is a variable created in CREATEV4.                            *0000000
*************************************************************************0000000
* Conditions:                          1 2                              0000000
*                                                                       0000000
C BFNCBU(I)=CLASS(BFNCBBI)            ;Y;ELSE;                         0000000
C BFNW_IU(I)=CLASS(BFNNONSP)          ;Y;                              0000000
C RGINDU(I)=CLASS(RGINDBI)            ;Y;                              0000000
C ASRU(I)=CLASS(ASRBI)                ;Y;                              0000000
C ETHIND(I) = 1                       ;Y;                              0000000
C ETHBLANK(I) = 1                     ;N;                              0000000
*                                                                       0000000
A BFNCBU(I)=D_YES_MEM                 ;X; ;                            0000000
A RGINDU(I)=D_YES_RGIND               ;X; ;                            0000000
A ASRU(I)=C_NAI                       ;X; ;                            0000000
A BFNW_IU(I)=A075_GET_SPEC            ;X; ;                            0000000
*                                                                       0000000
A RETURN                              ;X;X;                            0000000
```

Table 3: DLT After Translation to CANCEIS But Before Manual Clean-up

```
*
************************************************************************
* This table originates from ASRCKR.  It is the fourth table of the    *
* of the IBFN sub-module.  The tab checks if respondents provide any   *
* Aboriginal ethnic origin in Q17.  Derived variable, ETHIND, is used  *
* to identify Aboriginal ethnic origin.                                *
* ETHIND is a variable created in CREATEV4.                            *
************************************************************************
* Conditions:                          1 2
*
% DLT Name:                  ETHCKR

% Strata:
% Purpose:                  DERIVE
% Type:                     Conflict
% Symmetry:
% Sub-unit Start position:
% Sub-unit End position:

@ BFNCBU(#1)=CLASS(BFNCBBI)                        ;Y;
@ BFNW_IU(#1)=CLASS(BFNNONSP)                      ;Y;
@ RGINDU(#1)=CLASS(RGINDBI)                        ;Y;
@ ASRU(#1)=CLASS(ASRBI)                            ;Y;
@ ETHIND(#1) = 1                                   ;Y;
@ ETHBLANK(#1) = 1                                 ;N;
*
& BFNCBU(#1)=D_YES_MEM                             ;X;
& RGINDU(#1)=D_YES_RGIND                           ;X;
& ASRU(#1)=C_NAI                                   ;X;
& BFNW_IU(#1)=A075_GET_SPEC                        ;X;
```

16. A Windows Interface (called the Dictionary Importer) will be created to allow SPIDER Data Dictionaries to be imported (with some manual modifications) into CANCEIS.

17. Using the DLT Translator, the Dictionary Importer and the DLT Editor, Subject Matter personnel will be able to convert, with relative ease, the SPIDER DLTs to CANCEIS DLTs. Methodologists will help these personnel become familiar with CANCEIS plus the new Windows interfaces and assist with the conversions where necessary.

V. OTHER IMPROVEMENTS TO CANCEIS PLUS CONCLUSIONS

18. Besides the Windows interfaces, a number of other improvements are being made to CANCEIS. Features available in CANCEIS, but not used in previous censuses, are being tested to ensure that they work correctly. Error messages have been made clearer and are available in either English or French. Improvements have been made to an optional audit trail which shows exactly how CANCEIS determines which variables to impute. CANCEIS, when reordering persons in the failed household, will have the option of making it a higher priority to have the persons that enter the failing edits resemble those from the donor household. Improvements have been and will be made to the paper and on-line documentation. Other extensions to the CANCEIS methodology will be made, as required, to ensure that SPIDER modules are successfully ported to CANCEIS.

19. The significant challenge of porting all SPIDER modules to CANCEIS for the 2006 Census will be made easier because of the efforts to ensure similarity, where possible between the syntax of the SPIDER and CANCEIS. In addition, the use of Windows Interfaces to do these ports will reduce the amount of effort required.

References

Bankier, M. (1999), "Experience with the New Imputation Methodology Used in the 1996 Canadian Census with Extensions for Future Censuses", Proceedings of the UN/ECE Work Session on Statistical Data Editing, Italy (Rome). (http://www.unece.org/stats/documents/1999.06.sde.htm)

Fellegi, I.P. and Holt, D. (1976), "A Systematic Approach to Automatic Edit and Imputation", Journal of the American Statistical Association", March 1976, Volume 71, No. 353, 17-35.

Statistics Canada (2003). "Functional Description of the Generalized Edit and Imputation System". Statistics Canada Technical Report.