Using the New USDA Formats

Lois Steinfeldt, The University of Texas-Houston

HNIS has implemented new formats for the Survey Nutrient Data Base (NDB) System beginning with the 1994 survey. The new file structures were presented in detail at the 1993 NDB Conference and are in the handout entitled USDA SURVEY NUTRIENT DATABASE SYSTEM. The Survey NDB files include Survey and PDS descriptions, gram weights, measures and nutrient values; Survey recipes; retention factors and descriptions, moisture fat change values and descriptions and nutrient descriptions. They will be released as ASCII delimited files which have been exported from Paradox tables.

The Survey Food Codebook contains the information needed for coding foods and amounts. It consists of 7 files. The Codebook Description file contains food description data. The Include file contains "includes" which are similiar foods with comparable nutrient values and weights such as Wisconsin and New York cheddar cheese. Brand name foods can also be includes. The Codebook Subcode, Subcode Description and Subcode Include files contain information on subcodes. Subcodes are foods which have comparable nutrient values but different weights for the same measure. An example is Hostess cupcakes and Little Debbie Snak Cakes which are subcodes of the code for chocolate cupcake with icing or filling because of their similar nutrient profile. However, each of the brand name foods has a unique weight for the same measure. The PDS codebook files have the same formats as their comparable files in the Survey codebook.

The Gram Weights file contains a weight in grams for each measure description for a particular food item. Weights for similiar foods in the Survey and PDS codebooks are shared. The Measure Description file contains a 5 digit code for each unique measure description. For example 1 cup is measure number 10205.

The Survey and PDS Nutrient files contain nutrient values for the Survey and PDS foods respectively. Multiple nutrient values will exist when necessary to reflect changes that have occured in foods. Each nutrient value has a start date and an end date to mark the effective time period for that value.

The Survey Recipe files control the generation of Survey nutrient values using the PDS and other supporting files. There is a recipe header file and a recipe ingredient file. The recipe header file contains the recipe description and information on changes in moisture and/or fat that occur in cooking and the type of fat, if applicable. The recipe ingredient file contains information on recipe ingredients including the food code and description, amount and retention code.

Other support files include the Retention Factors files, Moisture and Fat Change files and nutrient description files. This is a brief overview of some of the information included in the handout. It lists each file structure and describes the relationships between the files and data items. Since successful use of data depends on an understanding of the data items and their relationship to one another, this is a good place to begin.

In addition to file format changes, there are changes in the size of some of the data items and the addition of new data items. The most notable change is the increase in the size of the Survey food code from 7 to 8 digits.

The new formats for codebook descriptions and weights and measures are much easier to access and manipulate than the previous text format. The ASCII delimited files can be imported into a variety of

software packages where the data can be queried and analyzed. Listed here are a few examples of how the data could be used.

SLIDE 1

Potential Uses for Survey and PDS Codebooks

- Create subset databases using food codes and/or descriptions
    Foods whose descriptions contain the word chicken
    Foods whose codes begin with 57 (cereals)

- Select and calculate gram weights for selected foods
    Average grams per surface inch of pizza

- Calculate nutrient values for standard measures for selected foods
    Sodium content for 1 cup of chicken soups
    Iron content of 3 oz raw weight of meats
    Vitamin A content for 1 cup of cold cereals

These are some examples of data records from the Survey NDB files in ASCII delimited format. The indentation indicates the continuation of a single record. In ASCII delimited files, the end of each data value is marked or limited by the use of a special character called a field delimiter. Since alphanumeric data items can contain the same character that is being used as a field delimiter, a different character is used to enclose alphanumeric fields. This is called a string delimiter. Common default delimiters are the comma as a field delimiter and the double quote as a string delimiter. Since commas and double quotes are used in food and measure descriptions, it is necessary to use a different character, the caret (^) symbol, as the field delimiter. The double quotation mark is the string delimiter for the alphanumeric fields. Dates are formatted as MM/DD/YYYY.

SLIDE 2

ASCII Delimited File Format Examples
Food and Measure Descriptions

- Codebook Description

11112000^   "Milk, cow's, fluid, lowfat, NS as to percent fat"^
             "MILK, COW'S, FLUID, LOWFAT, NS AS TO % FAT"^
             ""^"1"^4/01/1985^12/31/2010^

- Measure Description

61706^   "1 piece (1/8 of 7" x 12", approx 3-1/2" x 4")"^
          5/01/1993^12/31/2010^

There is no standard format for ASCII delimited files. Although software packages use basically the same specifications for reading and writing ASCII delimited files, there may be differences. Software packages may or may not provide the option of changing some or all of the specifications. The variation in ASCII delimited formats most often affects reading alphanumeric and date type fields correctly.

Specifications for the format of an ASCII delimited file include the character used as a field delimiter, the character used as the string or alphanumeric delimiter, and whether or not all fields or only alphanumeric fields are enclosed with the string delimiter. Software packages may allow the user to specify the format of the date field. This may be done as part of the ASCII delimited file specifications or part of the data base or software specifications. Most software packages will read and write a variety of date formats.

Using software defaults with alphanumeric fields which contain the default delimiters, may result in a loss of data. When the string delimiter occurs within an alphanumeric field, some software packages write the delimiter twice as a way of indicating it is not to be read as a string delimiter.

SLIDE 3

ASCII Delimited File Format Examples
Gram Weights and Nutrients

- Gram Weight

    11112000^0^0^1^10205^245^4/01/1985^12/31/2010^
    11112000^0^0^2^30000^30.6^4/01/1985^12/31/2010^

- Survey Nutrient

    11112000^203^3.339^4/01/1985^12/31/2010^
    11112000^394^.06^4/01/1985^12/31/2010^

When importing ASCII delimited files, some software packages scan the file to check the length for each field and the number of decimal places in numeric fields. The software package will then create a structure that matches the data. If the software does not, the structure must be created explicitly. In either case the field types, field widths and number of decimal places should be checked against the formats supplied by HNIS. Creating a structure based on viewing the first few records may result in incorrect formats and loss of data. As you can see in this example non-significant zeros and spaces are usually not included, it is not possible by looking at a few records to determine the maxiumum field size or maximum number of decimal places.

The most successful procedure for importing ASCII delimited files (or any other type of file) into software packages or using the files with custom programs includes these steps. First check the file and data format specifications. Second import a variety of test data to check for inconsistencies. Third review the data thoroughly after it has been imported and before you use it.

With the implementation of Trend Analysis in the Survey NDB, the starting and ending dates must be referenced in order to retrieve data accurately. In the Survey NDB files there is a start date and an end date for each record. When data is requested for a specific date a single record is returned. If no date or a range of dates is requested, decisions must be made about the meaning and processing of multiple records.

Slide 4 shows an example of a breakfast cereal which has food changes occurring on three different dates affecting five nutrients. The rest of the nutrient values remain the same. These changes reflect a combination of fortification and reformulation. The Vitamin A value, nutrient code 392, decreases on 04/01/1989 from 1324 to 794. The Vitamin C value, nutrient code 401, increases from 53.0 to 211.6 on 04/01/1987. The values for saturated fat, monounsaturated fat and polyunsaturated fat, nutrient codes 606, 645 and 646, change on 10/01/1989.

SLIDE 4

Survey Nutrient File
Multiple Food Changes to Nutrients

Survey Code	Nutrient Code	Nutrient Amount	Starting Date	Ending Date
57213000	392	1324.0	04/01/1985	03/31/1989
57213000	392	794.0	04/01/1989	12/31/2010
57213000	401	53.0	04/01/1985	03/31/1987
57213000	401	211.6	04/01/1987	12/31/2010
57213000	606	0.81	04/01/1985	09/30/1989
57213000	606	0.82	10/01/1989	12/31/2010
57213000	645	0.37	04/01/1985	09/30/1989
57213000	645	0.42	10/01/1989	12/31/2010
57213000	646	0.52	04/01/1985	09/30/1989
57213000	646	0.58	10/01/1989	12/31/2010

Slide 5 shows a simple selection command using the date of intake as the criteria used to select records. Using the Survey Nutrient file, records are selected for food code equal to 57213000; nutrient codes equal to 392 (Vitamin A), 401 (Vitamin C), 606 (Saturated Fat), 645 (Monounsaturated Fat) and 646 (Polyunsaturated Fat); and starting date less than or equal to the date of intake and ending date greater than or equal to the date of intake.

SLIDE 5

Survey Nutrient File
Data Selection Process

```
Use survey nutrient file

    Select the records with food code =  57213000  and

       nutrient code =  392 or 401 or 606 or 645 or 646  and

       starting date <= intake date and ending date >= intake date
```

Substituting in different dates for the intake date will retrieve different nutrient values as shown in Slide 6. In this example, none of the nutrients changed value more than once. However, that can and certainly will happen.

SLIDE 6

Survey Nutrient File
Data Selection Results

Date of		VIT A	VIT C	SAFA	MUFA	PUFA
Intake		392	401	606	645	646
03/15/1987		1324.0	53.0	0.81	0.37	0.52
03/15/1989		1324.0	211.6	0.81	0.37	0.52
06/15/1989		794.0	211.6	0.81	0.37	0.52
10/15/1989		794.0	211.6	0.82	0.42	0.58

Although the formats for the Survey NDB files have changed, the data contained in the files is basically the same. For example, the new Survey Nutrient file can be converted into the format currently in use. This will allow the Survey nutrient data to continue to be used with existing systems.

SLIDE 7

Survey Nutrient File Comparison

	Current		New	
Data Item	Format		Data Item	Format
----	---	-----	------	
Food Code	N 7		Survey Code	N 8
Food Description	A 51		Nutrient Code	N 5
Water	N 10.3		Nutrient Value	N 10.3
Food Energy	N 10.3		Start Date	Date
	.	.	End Date	Date
	.	.		
Potassium	N 10.3		SURVEY CODEBOOK DESCRIPTION	
			Survey Code	N 8
			Abbreviated Descriptor	A 60

The current format is fixed field. Each data item is in the same columns for all records. There is one record for each food with all the nutrient values in that record.

In the new format there is one record per food, per nutrient, per date range. Each record is identified by the food code, the nutrient code and the starting and ending dates. The size of the food code has been increased to 8 digits. The size of the nutrient values has not changed. The new Survey Nutrient file does not have the food description in it. All the food description data is now part of the Survey Codebook Description file. The food description data item is called the ABBREVIATED DESCRIPTOR and the size has been increased from 51 to 60 characters.

There are a number of ways to convert the nutrient file in the new format into the current format. It can be accomplished with a data base management or statistical package or with a programming language. Within a data base management or statistical package, the interactive commands or procedural language can be used. The two files needed are the Survey Codebook Description file and the Survey Nutrient file. In each of these files, records are selected for a specific date. The Survey Nutrient file which has one record per food per nutrient is converted into one record per food with all the nutrient values for that food in separate data fields. The two files are linked on Survey food code. In a programming or data base management procedure language it would look something like this.

SLIDE 8

Survey Nutrient File
Format Conversion Process

```
For each record in the Survey Codebook Description File
      where starting date <= selection date and
            ending date   >= selection date

      Save the food code and abbreviated descriptor in memory

      For each record in the Survey Nutrient File
         where food code = saved food code and
               starting date <= selection date and
               ending date   >= selection date

         Save the nutrient value in an array in memory

      Write food code, abbreviated descriptor and nutrient
         values to a record
```

A similiar type process can be used to created printed Survey and PDS codebooks. Each of the codebook files would be imported into a database management package. When a single version codebook is desired, selecting based on date as a first step will reduce the size of the files. The files can then be linked on the data items as described in the handout and a report created. Special purpose codebooks can be easily produced by selecting which foods and data items will be included and by varying the layout of the report.

Version data bases can be extracted for any date for any of all of the Survey NDB files. However the date used must always be a single date to insure that multiple values are not retrieved. For example, there are many ways a version data base could be set up for the calendar year 1994. The data could be selected based on the beginning, ending or middle of the year. Data could be selected based on the values in effect for the longest time period during the year or other more complex algorithms using weighted times.

In summary, the new formats for the Survey Nutrient Database System files will be easier to use and will make it possible to track changes in food data over time. The files are easily imported into software packages where they can be used for many different purposes.