Extracting data from a large instrument using the API

Lois Steinfeldt, ARS, USDA

1. Introduction

The Automated Multiple Pass Method (AMPM) instrument, developed by the Food Surveys Research Group, Agricultural Research Service, U.S. Department of Agriculture, collects 24-hour dietary recall data. The data are used to address economic, nutrition and food safety issues. For example, the data are used to evaluate the nutritional adequacy of the American diet and the impact of food assistance programs. The data are also used to estimate exposure to pesticide residues and to study the impact of food fortification, enrichment, and food labeling policies.

The steps in the AMPM interview are shown in Table 1. The multiple pass method improves the collection of 24-hour dietary recalls. Individuals recall the foods and beverages that were consumed the day before the interview. Details about each food and beverage are collected as well as an estimate of the amount consumed. Information is also collected on the time of day the food was eaten, the name of the eating occasion, whether the food was eaten at home or away from home, and where the food was obtained. AMPM contains more than 2500 questions and more than 20,000 responses. Ninety-five percent of the questions are about specific food details, including the amount of the food eaten.

Table 1. Automated Multiple Pass Method (AMPM)

	Pass	The respondent:
Step 1	Quick List	… reports an uninterrupted listing of all foods and beverages consumed in a 24-hour period the day before the interview.
Step 2	Forgotten Foods List	… answers a series of 9 food category questions probing for any further food items which may have been forgotten.
Step 3	Time and Occasion	…answers the time they began eating or drinking the food reported and what they would call the eating occasion for this food.
Step 4	Detail Cycle	…answers standardized questions developed by USDA to probe for detailed information about each food reported and the amount of the food eaten. Additional information is elicited about where the food or most of the ingredients was obtained and where each eating occasion was eaten. …reviews the eating occasions and times between occasions to see if additional foods are remembered.
Step 5	Final Review Probe	…answers a final probe for anything else consumed.

Foods are grouped into 132 categories each with a unique code. Each category has questions specific to the foods in the group. For example, a respondent reporting orange juice is asked if it was 100% juice and whether it was freshly squeezed, made from frozen concentrate, or came from a bottle, a carton, or a can. If the orange is from frozen concentrate, then the respondent is asked about the amount of water used to dilute the juice. When soda is reported, the respondent is asked if the soda contained caffeine and whether it was diet or regular. In addition to food details, AMPM provides the respondent with a number of ways to quantify the amount consumed. There are weight measures, volume measures such as cups and liters, and item descriptions such as 1 slice. In addition there are two-dimensional models of dishes such as glasses, mugs, and bowls, and shapes for measuring rectangular, round, and wedge-shaped foods.

The complexity and the size of the AMPM instrument make efficient data extraction and organization for food coding and analysis both a challenge and a necessity. In order to accomplish this, a program was needed which could quickly and accurately identify the fields with data values and extract the value and other selected field properties. Because of the flexibility and ease of programming, the Blaise Application Programming Interface (API) and Visual Basic were used to develop this program.

2. Background

There is a large amount of variation across respondents in both the numbers and the types of foods consumed. While the number of foods reported during an interview is usually less than 20, there have been a few respondents who have reported more than 30 foods. And while one respondent may have coffee many times during the day and no vegetables, someone else may have a lot of fruits and vegetables and no coffee. The AMPM instrument must balance the collection of complete and accurate intakes against the size and complexity of the data model. Setting array sizes to accommodate the greatest numbers of foods per day, and foods per category per day, produces a very large model. The limit for the number of foods per day is currently set at 40. So far this limit has not been exceeded. The number of foods per category per day is set at either 5 or 10 depending on how often foods in that category were usually reported. About one quarter of the 132 food categories accommodate up to 10 reports per day, the rest of the categories allow up to 5 reports per day. If the foods reported exceed any of the limits, the information is stored in a remark. Although the category limits have been exceeded a few times, for example with infant formulas, the limits in general have been adequate.

The AMPM data model contains information at the level of the interview, the food, and the food details. The majority of the data is stored in block and field arrays of foods and food details. For example, the food array, which allows for up to 40 foods, contains 53 fields including the name of the food, the time it was eaten, the name of the meal, and the food category. Because the food detail arrays contain the details and the amount consumed for the foods in that category, each is a different size. The milk category, which allows for up to 10 reports per day, has 33 fields including 5 of which record additions to the milk, such as chocolate syrup. The green salads category, which allows for up to 5 reports per day, has approximately 270 fields including the ingredients in the salad, the amounts of each ingredient and the type and amount of salad dressing.

Although there are over 140,000 defined data fields, as shown in the technical description of AMPM in Table 2, only 6,655 are elementary fields. Also shown in Table 2 there are 18,837 instances of 1,020 blocks. This shows that the size of the data model is due to the need to allow for multiple reports of foods and food ingredients, including their descriptions and amounts. For an individual interview, most of the fields in a record will be empty. From an average interview, there are fewer than 500 fields that contain data values needed for food coding and nutrient analysis. From an interview with more than 30 foods, there could be as many as 1000 fields needed, while from an interview for an infant there may be as few as 100. But because each respondent consumes different numbers and types of food, different data fields need to be extracted for each record. The difficulty lies in finding the data values that are needed without having to check every one of the

140,849 fields and in maintaining the link between the food detail data and other food information. Then once the data are extracted, they need to be organized for food coding and analysis.

Table 2. Technical description of AMPM - Overall Counts

	Value
Number of uniquely defined fields*1	7,675
Number of elementary fields*2	6,655
Number of defined data fields*3	140,849
Number of defined block fields*4	1,020
Number of defined blocks	1,020
Number of embedded blocks	234
Number of block instances	18,837
Number of key fields	7
Number of defined answer categories	1,923
Total length of string fields	239,2781
Total length of open fields	0
Total length of field texts	356,878
Total length of value texts	126,603
Number of stored signals and checks	127,397
Total number of signals and checks	127,397

*1) All the fields defined in the FIELDS section
*2) All the fields defined in the FIELDS section which are not of type BLOCK
*3) Number of fields in the data files (an array counts for more than one)
*4) Number of fields of type block

3. Data Extraction

Since less than 1% (~500/140,000) of the fields are going to be extracted, excluding large groups of fields as early as possible increases the efficiency of both the extraction and the subsequent processing. The main approach to extracting data is to use the information that is stored in the food array for each food reported to identify which food category array and which instance in the food category array contains the food detail information for that food. When a respondent reports a food, it is selected from a food list. Each food in the list is linked to a food category that determines which questions are asked for the food. Information about the food is stored in a food block array. Because the same category of food (e.g. fruits) can be eaten more than once a day, the answers to the food detail questions are stored in arrays for each food category block. Both the food category and the instance in the food category block array are stored in the food block array for each food reported. Using these fields, the extraction program can find and read only one food category detail block instance for each food. This greatly reduces the number of fields in the AMPM database that must be read.

Once the correct block instance is found, the extraction uses the basic recursive function ‘For Each objField in ParentField Fields’ expanded to reference a set of exclusion criteria. The exclusion criteria were added because even limiting the extraction to the specific food category block instance, the program would still extract many fields that aren’t needed. For example, the meat sandwich category must allow the respondent to report multiple meats, cheeses, and vegetables in a sandwich. Here again, the variation in food consumptions across individuals requires that enough space be allocated to record the sandwich with three meats, two cheeses, and five vegetables, as well as the sandwich with one meat, one cheese, and no vegetables. For most sandwiches this results in a lot of empty fields. Although the fields must be accessed to determine if they meet the exclusion criteria, subsequent steps in the processing benefit by not extracting the fields that are not needed. The exclusion criteria used are collections of Field Tags, Field Values, and Field Names, which are evaluated in that order. The order is based on the likelihood that the criteria will exclude a field or a group of fields. That likelihood is based upon the AMPM data model and the nature of the data collected. The number of items in each of these collections is kept as small as possible to reduce the amount of time the program spends comparing the field properties in the database to the items in the exclusion collections.

The exclusion criteria used most often are Field Tags and Field Values. The consistent naming and use of Field Tags in AMPM made possible this quick and efficient method for eliminating fields that contain fills, which have a Tag Name of INTFILL and fields that are used to control and monitor the flow of the instrument which have a Tag Name of INTFLOW. Then because such a large percentage of the fields are empty, exclusion based on empty Field Values greatly reduces the amount of data extracted. Field Names are the least utilized because of the amount of work to specify and maintain a Field Name collection and the time it would take the program to compare each Field Name to a large collection of Field Names. For AMPM, the Field Name collection contains a single entry that occurs in a large number of the food category block arrays and which cannot be eliminated using any other criteria. This field contains the answer to the question, “Did you add anything to this food?” While the subsequent food coding process does need to know what foods were added, it does not need to see the yes or no answer.

Before the exclusion criteria are evaluated, the presence of a remark is checked. There are a few fields in AMPM where the field can be empty, but the interviewer is able to record a remark. If there is a remark attached to a field, that overrides the exclusion criteria and the field is extracted.

In addition to the exclusion criteria listed above, there are additional criteria used at the data model level and at the first and second block levels. They are Include Root level fields (yes or no) and a collection of Block Names. The Root level fields criteria either includes or excludes the fields at the level of the object passed to ‘For Each objField in ParentField Fields’. The block name collection allows the exclusion of blocks that are used to control the flow of the instrument and do not store any food information. It is also used to exclude blocks that only contain interview instruction fields.

During extraction, each field is written into a Microsoft Access database as a separate record with a unique identifier. The unique identifier is a sequential number representing the order the field was extracted from the Blaise databases. This number is important for keeping the food detail fields in order, which is needed for accurate food coding. The Blaise database name and primary keys further identify each record. Since AMPM contains food level data in addition to interview level data, the food number further identifies the food specific fields. Each record also contains the fully qualified name, local name, display text, value, type, tag, and remark for the field.

4. Conclusion

The extraction program developed with the Blaise API and Visual Basic successfully identifies and extracts the intake and food data values that are required from each record. Parameters were created to allow data fields to be excluded from the extraction based on block name, tag name, field name, and/or field value. Once the blocks that contain food detail data are identified, the program quickly steps through the fields within each block testing for exclusion criteria. The presence of a remark, even on an empty data field, overrides the exclusion criteria so that all fields with remarks are extracted. The program has proved to be an efficient and accurate method to extract the small amounts of intake and food data from the large AMPM database.