Nutrient Data Bank: Computer-Based Management of Nutrient Values in Foods

RITVA R. BUTRUM and **SUSAN E. GEBHARDT**, Consumer and Food Economics Institute, ARS, USDA, Hyattsville, Maryland 20782

ABSTRACT

A computerized Nutrient Data Bank has been designed for storage, summary, and retrieval of food composition data. The system is a repository for data from domestic and international sources, including research institutions, industry, and independent laboratories. Source data are carefully screened with regard to identification of the food and conditions which may affect its nutritive value. Variables such as treatment and processing of the food and method of nutrient analysis can be considered in the analysis and retrieval of the data. All primary data will go into Data Base I. After statistical analysis of primary data, unique criteria will be developed for each food for use in summarizing the nutrient data into composite values. Data Bases II and III will be derived from the information in Data Base I by averaging, weighting, and selection. The summarized data will include averages for each nutrient, the number of samples, range values, and standard error. The data can be used for compiling a new nutrition handbook and for rapid retrieval of information for scientists.

INTRODUCTION

Food composition tables are valuable tools in the field of nutrition not only for planning nutritionally adequate diets but also for evaluating food supplies and planning programs for food distribution. Their importance has been recognized for over three-quarters of a century, and extensive amounts of food composition data have been published during that time. One of the most comprehensive of these tables is the "Composition of Foods...Raw, Processed, Prepared" published by the USDA (1).

Today, there is increasing need for revising and updating of food composition tables because of the greatly increased volume of data (2,3). Interest in food composition has been stimulated by the critical shortage of world food supplies, both in quantity and quality; by nutritional studies indicating possible nutritional deficiencies, even in populations with adequate food supplies; and by nutrient labeling, both voluntary and required, of foods in the United States and other countries. Much of the recent analytical work is on new foods never before included in food composition tables and on previously included foods analyzed by improved analytical techniques which are considered more reliable than older methods and have therefore replaced them. Ideally, as new analytical information on foods becomes available, it should be incorporated promptly into food composition tables.

The collection of nutrient values and their incorporation into meaningful food composition tables are painstaking tasks which require meticulous care on the part of the compiler. This work entails not only searching for the data but also a critical review of the data to categorize and classify the analyzed food items and to determine whether appropriate analytical methods were employed.

In order to free scientists from time consuming hand calculations, automated biostatical computing methods are urgently needed. With such methods not only could calculations be made in a fraction of the time required by present methods but also other computations, not now feasible, could be performed. Moreover, data could then be tested on different parameters, alone or together, to determine more rapidly and efficiently the effects of such factors as growing conditions or storage times.

USDA, with cooperation from the Food and Drug Administration, the Canadian Government, and the food industry, have developed a computerized Nutrient Data Bank (NDB) for storage, summary, and retrieval of analytical data on foods for both domestic and international use. This Nutrient Data Bank will serve as an efficient means for developing and updating food composition tables. This paper discusses the methods for obtaining data, the classification and categorizing of food items, the coding system for data entry, and the evaluation and summary of analytical data on foods within the computerized system.

DATA ACQUISITION

A primary function of the NDB is to be a repository of all data on the nutrient composition of foods. However, nutrient data are available from a broad spectrum of sources ranging from chemical journals and state experiment station bulletins to food industry laboratory reports. The success of the system depends on the acquisition of nutrient data from all of these varied sources.

The Current Awareness Literature Search of the National Agricultural Library is used to obtain published nutrient data (4). This system is a computer-based search of pertinent scientific data bases, Biological Abstracts, Chemical Abstracts, Food Science and Technology Abstracts, and the National Agricultural Library Catalog (CAIN).

To achieve maximum retrieval, the search for food composition data must be written without too many limiting parameters. This results in retrieval of practically all pertinent documents, but also of many irrelevant citations. Sometimes it is possible to eliminate useless articles by their titles, but often it is necessary to order and review an article to realize that it does not contain useful data—a time consuming procedure.

Present food tables lack data on new processed food products. The literature reveals few articles on the nutrient composition of these products, and correction of this problem depends primarily on the willingness of companies to submit nutrient information about their products to the NDB. Many companies are having their products analyzed for nutrition labeling. Label declarations as percentage of the US Recommended Daily Allowances (RDAs) are not useful for the NDB because actual values or average values with some indication of variability are required. Laboratory reports from which the percent US RDAs are calculated are ideal sources of data. Data may be submitted in any form most convenient for the supplier of the data—laboratory reports, computer printouts, or tapes. An adequate description of the food item should accompany the nutrient information.

Incoming data are assigned to a nutritionist or chemist at the Institute for review before being coded for entry into the system. The food specialist checks for obvious errors in the data presentation; sufficiency of identification of the food; and adequacy of information on processing, sampling, and analytical procedures. In situations where the data are questionable but no further clarification can be obtained, the data will be flagged. A flag will prevent these values from being included in any summaries but can be removed

TABLE I

MAJOR FOOD GROUPS

A.	MEAT (other than poultry)
B.	DAIRY PRODUCTS
C.	POULTRY, REPTILES, THEIR EGGS; and INSECTS
D.	FISH, SHELLFISH, THEIR EGGS and PRODUCTS
E & F.	MATURE LEGUMINOUS SEEDS, NUTS, and OTHER SEEDS
G.	SAUSAGES and LUNCHEON MEATS
H & I.	MIXED DISHES with ANIMAL PROTEIN
J.	SOUPS, BROTHS, MEAT EXTRACTS, and NATURAL MEAT JUICES
K.	PROCESSED MEALS, MEAL REPLACEMENTS, SANDWICHES, HOT DOGS, HAMBURGERS, and OTHER BUN-TYPE SANDWICHES
L.	CEREAL GRAINS and THEIR MILLED PRODUCTS
M.	BAKERY PRODUCTS
N.	VEGETABLES and VEGETABLE PRODUCTS
O.	FRUITS and FRUIT PRODUCTS
P.	MIXED DISHES, DESSERT TYPE (other than bakery and fruit products)
Q.	SUGARS, SYRUPS, ICINGS, BAKING CHOCOLATES, and COCOA
R.	PROCESSED FATS, OILS, and SALAD DRESSINGS
S.	BEVERAGES
T.	SEASONINGS and CONDIMENTS
U.	LEAVENING AGENTS and ADDITIVES
V.	BABY FOODS
W.	SIMULATED PRODUCTS (fabricated foods)

later if the values are found to be accurate. Information that is to be coded about the product for entry into the system is indicated by the specialist.

FOOD NOMENCLATURE

As the name of the food is the first data element coded, accurate identification of a food item is naturally imperative. Confusion is often created by lack of standardization in nomenclature. A food item for which nutrient data are published may be known to the author by another name than that used by the reader. Therefore, name lists of primary and processed foods have been developed. The list of primary foods includes commonly used names and scientific names for food items. All common names for a particular food are cross referenced to indicate the preferred or universal name. This list of common and scientific names will be kept up to data as the program develops.

Standarized nomenclature is also used for processed foods. Because brand names are not used in the system, the items are assigned a descriptive term which will encompass a group of similar food items.

Because of the numerous formulated foods and processes used in the preparation of foods for consumption, there is a need for identification of the terms used to describe these foods and processes. A glossary of food terms has been established which gives definitions for processes and treatments, parts of foods, stages of maturity, and other terms that may need clarification. The use of a unified terminology lends greater validity to the system, makes the system easier to use, and eliminates confusion caused by large numbers of names and food terms used interchangeably.

FOOD CLASSIFICATION AND IDENTIFICATION

Each entry into the data bank will have a food identification code consisting of two parts. The first part is a four-position alphabet code called the food term with each position corresponding to a specific level in a four-level food classification hierarchy. Twenty-one major food groups (Table I) form the highest level in the hierarchy. These major groups are divided into major subgroups at the second level, and the major subgroups are divided into minor subgroups at the third level of the hierarchy. Individual foods are designated at level four.

All items within each level have been assigned specific alphabet letters to represent them within the food term code. For example, the code "BLAI" is the food term for blueberry yogurt. B represents the first level or major group which is dairy products (Table I); L, the second level—cultured dairy products other than milk or cream; A, the third level—yogurt; and I, the fourth level (that is, the actual food item)—blueberry yogurt.

The second part of the food identification code is used to describe further or modify the food term. It is made up of "qualifying" terms (Table II). While each entry into the data bank can have only one food term, it may have up to 32 qualifying terms. These include factors that may be important in evaluating the data, such as type of preserving technique or storage condition. A list of terms that may be used in this capacity has been compiled, arranged into categories, and assigned codes. The first position of the code is an alphabet character and designates the particular category of qualifying terms. Sequential three-digit numbers are used in the remaining positions.

TABLE II

Categories of Qualifying Terms

QUALIFYING TERMS

A.	**TREATMENT APPLIED**
B.	**PRESERVING TECHNIQUE**
C.	**PROCESSING TECHNIQUE**
D.	**COOKING METHOD**
E.	**PHYSICAL STATE**
F.	**PORTION ANALYZED**
G.	**PACKAGING AND STORAGE CONDITIONS**
H.	**GRADE, QUALITY, APPEARANCE, SIZE AND COLOR**
I.	**MATURITY AND CONDITIONS OF GROWTH AND PRODUCTION**
J.	**SPECIAL DESCRIPTORS**
L.	**CATEGORY OR VARIETAL TYPE**
N-R.	**COMPONENTS OF MIXED DISHES**

An example of how the coding system will be used to include qualifying terms is shown in Figure 1 where codes have been developed for frozen, sliced, California grown Elberta peaches, with sugar and vitamin C added. The food term is "OBKF," with O representing the major food group—fruits; B, the subgroup—stone fruits; K—peach; and F, the variety—Elberta. The other codes represent qualifying terms which designate the preserving technique—frozen; physical state—sliced; production location—California; and added components—sugar and vitamin C.

The structure of the four-position alphabet code for the food term and the alpha-numeric code for the descriptive or qualifying terms offers numerous possibilities for codes. At present, approximately 10,000 food terms and 3,000 qualifying terms have been identified, and both parts of the coding system readily allow for expansion.

FIXED DATA ELEMENTS

Information may be recorded for over 200 different nutrients and other items relative to food composition, including all nutrients presently recognized as essential, as well as ash, cholesterol, protein efficiency ratio (PER), pH, and total solids. A unique three-digit code identifies each

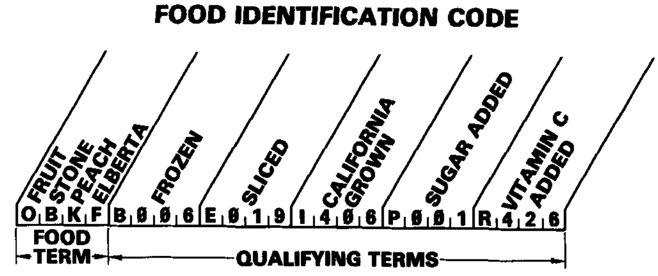

FIG. 1. Food Identification Code

nutrient. For each nutrient for which data are entered, the method of analysis is coded and entered. The method code has been structured to indicate necessary details of the method such as extraction procedure or type of hydrolysis as well as any modification which has been made to a standard method.

In addition to the food identification code and nutrient information, there are other items about a product which are coded and entered into the system for summary and retrieval purposes. Each data source, whether a journal article, computer tape or printout, personal communication, or data collection form will be assigned a *document* number. Within a document there may be data for several different food items or just one food which has received various treatments. Each item will receive a different subset number. For example, one document might contain data on the following foods: green beans, canned; green beans, frozen; peas, canned; and broccoli, frozen. These items would receive subset numbers 0001, 0002, 0003, and 0004 respectively. The document number and subset number will specifically identify each item for which data are entered within the system.

Published data will be coded to indicate the source according to the CODEN system (5). This system, prepared and maintained by the Science Information Services, provides an abbreviated title for journals and other published documents from which data might be obtained. Included with the CODEN abbreviation will be the year, volume, issue, and page number; these will provide a complete source identification. Unpublished data will receive a reference code to identify the data as being from industry, state experiment station, government, or university.

Geographic production area will be recorded for each entry of data whenever possible. The Federal Information Processing Standards (FIPS), established by the U.S. Department of Commerce-National Bureau of Standards, have been adopted to code countries and the 50 US states (6). Where feasible, provinces or regions within other countries as well as all major marine areas will be coded.

Unit conversion has been built into the system. All data for a given nutrient will be stored within the system in a common or standard unit. For example, preformed vitamin A reported and entered in International Units will be converted to milligrams of retinol. This feature should greatly facilitate the entering of data into the data bank.

ENTRY AND SUMMARY OF THE FOOD COMPOSITION DATA

After all necessary codes are assigned to an analysis, the codes along with the nutrient values are transferred to machine readable form and entered into the system. The first program in the system, the update screening and validation program, will check for proper input. It will verify all codes and make cross reference checks with builtin parameters for nutrient values. For example, it will indicate values which are above or below values that can be reasonably expected. If a unit conversion is needed, it will be performed at this time.

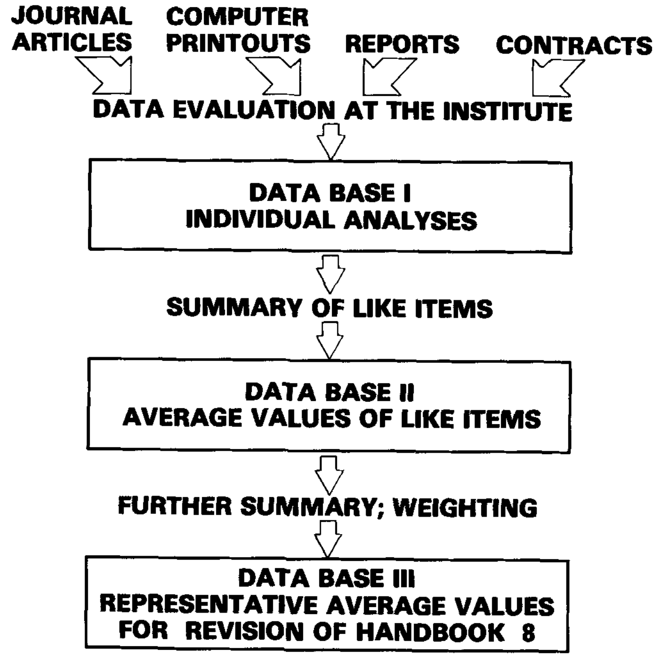

FIG. 2. Flow of data through Nutrient Data Bank system

All the individual data with detailed information will now go into Data Base I (Figure 2). Data Base II will be derived from Data Base I and will contain averaged analyses of identical food items—that is, all food items with matching four digit food codes and matching qualifiers. In addition, some food items which are similar but not identical may be summarized together if there is no significant difference in nutrient content. However, some nutrients in some foods, such as the vitamin C content of Valencia oranges grown in Florida and California, may be statistically different. If so, these items would not be summarized together and Data Base II would contain nutrient information for both items. Where data permit, statistical tests will be performed to determine what factors influence the nutrient content of foods and where significant differences occur. Data Base II will also include the standard error as well as the number of samples and the range of values which comprise the mean. Data Base II is expected to be quite extensive and further summary will take place to build Data Base III. In addition, items will be weighted to reflect food production figures for data Base III.

Data Base III will be equivalent to a food composition table. It will contain over 4,000 food items and its mean values will be the same as those published as revised USDA Handbook No. 8, although it will list more nutrients than will the published version. Generally, one item representative of each food consumed in the United States will be present, but, when data warrant, values for important subgroups will be included. For example, there may be only one item for flatfish, in which nutrient values for flounder, sole, and dab have been summarized together. While this same type of summary may be performed for a fruit such as apples the most common varieties or cultivars of apples may remain as individual items. The number of samples, the range of values, and the standard error will also be present.

Data Bases II and III will be contracted to a private company where they will be available on-line for anyone wishing to access the data. Also, magnetic tapes of these data bases may be purchased. Presently, Data Base I is being built and this on-line system is not available. The Dairy Products and Egg section will be the first group to be made available.

The Nutrient Data Bank with its computerized system for storage and retrieval of data will greatly increase the present capacity for processing large amounts of food composition data. Discussions have been held in regard to the exchange of information between the Nutrient Data Bank and the Food Composition Data Management System of the Food and Agriculture Organization (FAO). It is hoped that cooperation between FAO and USDA would produce more reliable nutrient data on a global scale with a minimum of duplication of work. With sufficient enlargement of the data base, the Nutrient Data Bank will serve as an international source of data for different regions of the world.

REFERENCES

1. Watt, B.K., and A.L. Merrill, "Composition of Foods—Raw, Processed, Prepared," USDA Handbook No. 8, 1963.
2. Murphy, E.W., B.K. Watt, and R.L. Rizek, J. Ass. Offic. Anal. Chem 57:1198 (1974).
3. Watt, B.K., S.E. Gebhardt, E.W. Murphy, and R.R. Butrum, J. Am. Diet. Ass. 64:257 (1974).
4. Current Awareness Literature Service, "User's Guide 1974," 2nd Edition, Data Systems Application Division, ARS, USDA, 1974.
5. Blumenthal, J.G., Ed., "CODEN—for periodical titles," American Society for Testing and Material-Science Information Services, Franklin Institute Research Laboratories, 1970.
6. U.S. Department of Commerce/National Bureau of Standards, "Federal Information Processing Standards," FIPS publication 10-1, 1974.

[Received June 21, 1976]

Pittsburgh topics announced

The 1977 Pittsburgh Conference on Analytical Chemistry and Applied Spectroscopy will be held Feb. 28 to March 4, 1977, in the Cleveland Conference Center, Cleveland, OH.

Fourteen symposia are scheduled including "Chromatography as Applied to the Food Industry," arranged by P. Talarico; "Characterization, Quantitative Determination, and Control of Odors," arranged by Dr. R.W. Freedman; "ASTM-31: Computerized Laboratory Systems," arranged by Dr. J.W. Frazer and Dr. G.W. Barton; and "Analytical Techniques on the Horizon," arranged by W.M. Hickman.

The conference is sponsored by the Society for Analytical Chemists of Pittsburgh and the Spectroscopy Society of Pittsburgh. Registration material is available from Dr. A.J. Sharkins, Alcoa Laboratories, Alcoa Center, PA 15069. Housing information is available from Dr. J. Kevin Scanlon, Pittsburgh Plate & Glass, P.O. Box 11472, Pittsburgh, PA 15238.