Journal of Food Composition and Analysis 22 (2009) 433–441

Contents lists available at ScienceDirect

Journal of Food Composition and Analysis

journal homepage: www.elsevier.com/locate/jfca

Original Article

USDA's Nutrient Databank System – A tool for handling data from diverse sources

D.B. Haytowitz *, L.E. Lemar, P.R. Pehrsson

Nutrient Data Laboratory (NDL), Beltsville Human Nutrition Research Center, USDA-ARS, 10300 Baltimore Ave., Beltsville, MD 20705, USA

ARTICLE INFO

Article history:
Received 10 January 2008
Received in revised form 24 June 2008
Accepted 13 January 2009

Keywords:
Food composition
Database
Nutrient data
Food item table
Nutrient value table
Data compilation
USDA Nutrient Databank System
NDBS
Nutrient Data Laboratory
NDL
Dr. Atwater
What We Eat in America
NHANES

ABSTRACT

Key features of USDA's Nutrient Databank System (NDBS) allow processing of food composition data from diverse sources, including USDA's National Food and Nutrient Analysis Program, the food industry, scientific literature, and food labels. The Nutrient Data Laboratory (NDL) designed the NDBS as a three-tiered ("Initial", "Aggregation", and "Compiled") data management system to facilitate handling of data. Raw data and documentation (data source, sample description, sample handling, and analytical methods) are migrated into the Initial module. NDL scientists compare new data with old values and decide how to combine the initial data into aggregated data. In the Aggregation module, data can be grouped and weighted by parameters such as study, source, and market share. Depending on the type of data, various statistical algorithms are used to generate statistics, such as mean, standard error, number of data points, and error bounds. In the Compiled module, food names are finalized and common measures selected. Nutrient profiles are developed and missing nutrients/food components are imputed according to standardized scientific principles. A formulation application employing linear programming techniques, estimate, formulations for commercial foods and nutrient profiles based on the nutrient content of ingredients and target values derived from label information. A recipe application calculates nutrient profiles based on ingredients and their known proportions, allowing for the application of food yield and nutrient retention factors. The NDBS automatically documents how each value was derived and incorporates quality control checks at all levels. Prior to release, the completed nutrient profiles are reviewed by NDL scientists and, if approved, disseminated. The NDBS brings together a number of stand-alone modules and applications into one integrated system allowing the management of ~7500 food items for up to 140 nutrients/food components. Data points and documentation are managed and maintained in one place, providing an "audit trail" for each data point. The NDBS contains algorithms to assign confidence codes using NDL's data quality evaluation system. The NDBS permits the annual release of reliable data for a comprehensive set of nutrients/food components for a wide variety of foods on NDL's Web site: http://www.ars.usda.gov/nutrientdata. Through these releases, NDL provides food composition data for researchers, diet and health professionals, and consumers, including the "What We Eat in America" component of the National Health and Nutrition Examination Survey (NHANES).

Published by Elsevier Inc.

1. Introduction and historical background

The U.S. Department of Agriculture (USDA) has maintained tables of food composition for over 115 years, since the pioneering work of Atwater (Atwater and Woods, 1892). Published in 1892, the Atwater table of 178 food items contained data on five proximate components (water, protein, fat, total carbohydrates and ash), kilocalories (called fuel in the Atwater table), and refuse. Dr. Atwater's data sheets (Fig. 1) are currently maintained in the Special Collections Section of the National Agricultural Library in Beltsville, MD (http://www.nal.usda.gov/speccoll).

The historic tables are similar in content to the electronic spreadsheets used today for many aspects of the work. This work is currently conducted in Beltsville, MD by the Nutrient Data Laboratory (NDL), which is part of the Beltsville Human Nutrition Research Center of the Agricultural Research Service of the USDA.

With the expansion of the number of nutrients/food components included in the tables, the increased quantity of data, and the need to capture more information about the food samples, it became clear that computerizing the data would facilitate the process of compiling tables of food composition. The first Nutrient Databank System (NDBS) ran on the USDA's mainframe computer and was written in the COBOL programming language (Table 1).

In the 1980s, the system was redesigned and written in a new programming language, PL1. Over the next decade, the NDL developed a number of stand-alone applications to allow

* Corresponding author. Tel.: +1 301 504 0714; fax: +1 301 504 0713.
E-mail address: david.haytowitz@ars.usda.gov (D.B. Haytowitz).

0889-1575/$ – see front matter. Published by Elsevier Inc.
doi:10.1016/j.jfca.2009.01.003

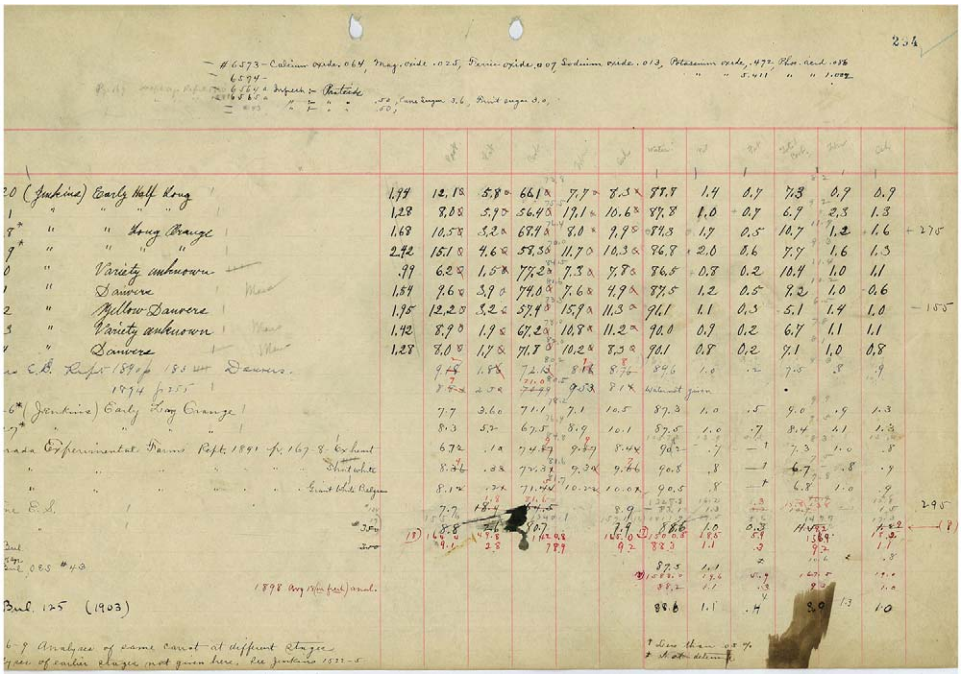

Fig. 1. Sample of Dr. Atwater's worksheets.

expansion of its database and to accommodate new data needs. For example, shortly after the completion of the 1980s databank system, a formulation estimation application was developed using the General Algebraic Modeling System to perform linear programming calculations (Marcoe and Haytowitz, 1993). This stand-alone application enabled nutrient estimates to be calculated by recipe using estimated ingredient percentages when analytical data were not available. This application was invaluable in allowing NDL to expand data for emerging nutrients of public health significance to be used in food intake surveys and epidemiological work.

By the mid-1990s, the need for a totally redesigned system was recognized. This report describes features of the latest version of the USDA's NDBS, the Architecture and Integration Management: Nutrient Databank System (AIM_NDBS). Though this system has existed in two architectures, one Client-Server (Oracle Client) and the other Web-based (Oracle Web), the features and functionality are the same.

2. System overview

The goal of the AIM_NDBS development was to provide an integrated system to handle acquisition, evaluation, compilation, storage, and dissemination of food composition data. The NDBS was developed to integrate all critical staff functions associated with nutrient data work, particularly those functions associated with development and dissemination of the USDA National Nutrient Database for Standard Reference (SR) (NDL, 2007a). Planning for the system began in 1997.

Table 1
Timeline and computer applications used in developing the USDA Nutrient Databank System.

Year	System	Platform and software
1976	NDBS	Mainframe – COBOL
1980	NDBS	Mainframe – PL1
1990	Master Database	Corel® Paradox/® – Novell® Server
1997	AIM_NDBS	Windows NT Server – Oracle® Client-Server
2005	AIM_NDBS	Windows NT – Oracle® Web-based application

The project to develop the new system was divided into three major phases:

Phase 1: Determination of system requirements.
Phase 2: Hardware and software selection.
Phase 3: Final system design, development, procurement, testing, and installation.

The AIM_NDBS utilizes a relational database comprised of 271 tables containing various sets of data to support the system, residing in an Oracle® database.

The NDBS was designed as a three-tiered data management system, comprising three modules:

1. The Initial Food Item module – Detailed nutrient/food component, weight, and physical component (i.e., part of plant or animal determined by dissection, including flesh, peels, bones, etc.) values are entered and food item description and methodology information are documented.
2. The Aggregation Food Item module – Individual nutrient/food component, weight, and physical component data for similar food items can be aggregated.
3. The Compiled Food Item module – Missing nutrient/food component values are imputed using standardized procedures, including recipes and/or formulations, and the food item profile is finalized for dissemination.

In addition, the NDBS has a variety of support areas which can be categorized as:

1. Utilities – Applications that assist NDL in maintaining and updating a variety of support tables; provide factors necessary for many unit conversions; and provide functionality for rating data quality, calculating nutrient retention factors and estimated cooking yields, and disseminating several databases, including SR.

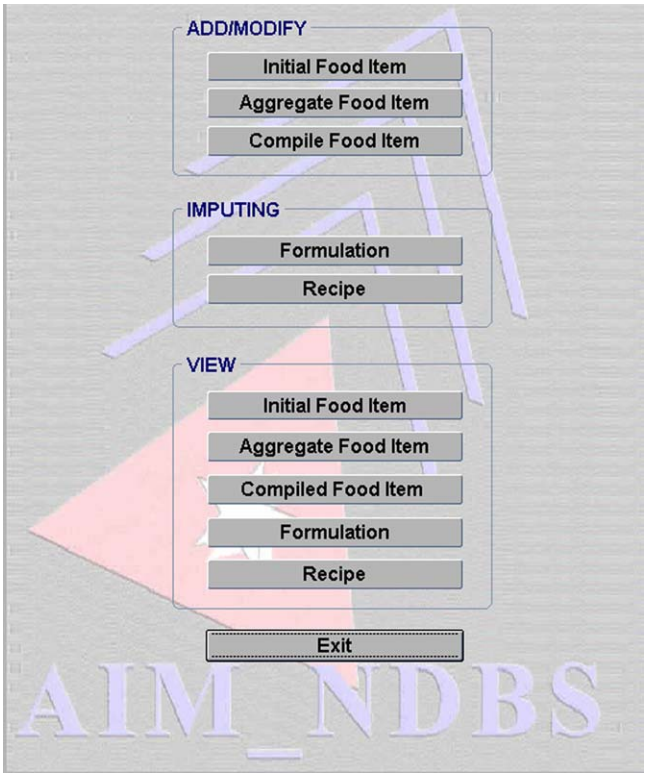

Fig. 2. Screen of the NDBS basic menu.

2. Reports – Over 30 pre-defined reports can be generated for viewing and printing. These reports include: (1) content reports for a particular food item at any level of the system; (2) auditing reports which track data processing throughout the system; and (3) reports comparing differences in nutrient data between two related food items.

Fig. 2 displays the main menu with its separate edit (add/modify) and view modes, access buttons for the three tiers of data processing (Initial, Aggregation, and Compiled modules), and direct access to two major imputing applications (formulation and recipe). Context-sensitive help is available throughout the system.

A comprehensive set of quality control/data validation tests is built into the system. This complements NDL's quality control (QC) program where data originating from USDA contractors are reviewed before they are migrated into the system. At each level, specific QC procedures must be run, and the item approved for release before it can be used at the next level. QC procedures are customized for each level and perform basic data validation steps (e.g., the sum of the proximates per 100 g of food should not exceed 100 g). At the same time, various nutrients are also calculated. For example, in the Compiled module, the energy (kilocalories and kilojoules) and "carbohydrate by difference" (100 minus the sum of moisture, protein, fat, and ash) values are calculated.

2.1. Initial food item module: data entry and storage of raw data

Before AIM_NDBS was developed, a large amount of documentation, now stored electronically in the databank, was kept on paper. For example, analytical methods were assigned code numbers that were entered into the databank system, but the method definitions and references were stored on handwritten or typed index cards. Food labels were often photocopied and stored in folders, but there was no way to capture or directly use most of the nutrient or ingredient information in the older (legacy) databank systems. Likewise, documentation needed for data quality evaluation using expert system procedures was not integrated into the system. One of the major operations in implementing the new system was converting previously generated data and documentation to the new system and determining how these data fit into the newly created database model.

In the first tier of the databank system, the Initial module, all of the individual data points are maintained. Thus, percent Daily Value (DV) information (reference values for nutrients used in U.S. food labeling) can be stored; analytical data in the exact units as received (e.g., fatty acids as percent methyl esters) can also be stored. At the end of the Initial module processing, all data will be converted to standard units per 100 g of food. Other informations documented for analytical values include: method of analysis, analytical quality control, sampling plan, and sample handling. This information can be used to generate data quality ratings using the procedures described by Holden et al. (2002, 2005). Work is ongoing to provide data quality measures for all nutrient/food component values. Common measures, physical components such as refuse data, as well as the procurement source and sampling information for each individual food sample, are also maintained in the system. For many food samples and for many brand name items provided by the food industry, label information such as ingredients, data from the Nutrition Facts panel, preparation instructions, and the UPC code are also entered.

The system contains a number of scripts (short computer programs) for importing the various types of data. In addition, data may be entered manually. As appropriate, food yield information for various cooking methods is collected and entered to enable calculation of yield and nutrient retention factors through the use of a specific utility.

The Initial form is divided into three parts:

1. Menu bar – Provides various functions, such as adding a new item, querying the database for a specific initial item or group of items, determining the impact of changes to this record on other records, associating different types of initial data for a single food item, and returning to the "main" menu.
2. Header – Gives general information on the food (source of data, detailed food description, manufacturer, research sponsor, e.g., USDA, another government agency, food company, or trade associations).
3. Specific tabs – Nine "tabs" (Table 2) containing various categories of information are available for each food item in the Initial module. Fig. 3 shows the nutrient data tab for broccoli, as an example.

Table 2
Description of specific tabs available within the Initial module.

Tab	Description
Sampling plan	Information on the specific sampling plan that was developed and used, and where the sample was obtained
Handling	Information on how the sample was stored, homogenized, and composited
Label	Ingredient information, UPC codes, and expiration dates
Name	Alternative names for the product as well as scientific names
Weight	Data on the gram weight of various food-specific common measures
Nutrients	Nutrient values and related information, such as method of analysis, laboratory performing the analysis, quality control, source and derivation codes, and statistical data
Physical components	Data on the various parts of the food item (e.g., bones, seeds, peels, etc.), and any identified as refuse
Preparation method	Data on the preparation methods, including heating processes, e.g., cooling, thawing
LanguaL	LanguaL terms for describing the food item (EuroFIR, 2007)

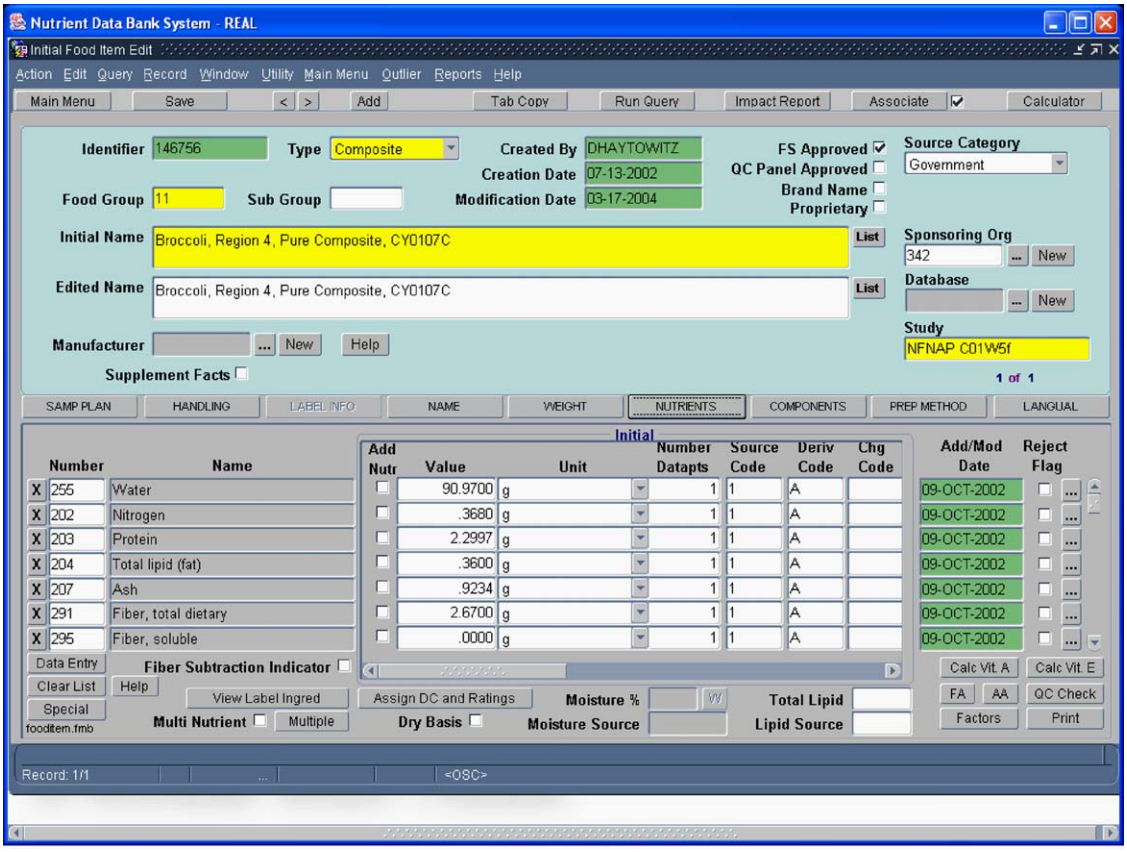

Fig. 3. Sample of initial data: nutrient data tab.

An ambitious analytical plan, the National Food and Nutrient Analysis Program (NFNAP), was initiated in 1997 to update and improve data in the USDA National Nutrient Databank and the databases disseminated from the system (Haytowitz et al., 2007). Planning for both the NFNAP and the redesign of the databank system occurred concurrently, allowing for the coordination of system development with NFNAP protocols.

A major challenge when designing the Initial module was associating all data belonging to a particular food item sampled and analyzed under the NFNAP protocol. The solution was to incorporate an "Associate" function within the Initial module to combine the three types of records common to NFNAP items, namely sample units, subsamples, and composites (Fig. 4).

Sample Units – These are the individual items or units that are procured or selected in accordance with the sampling plan developed for the particular food item. The information documented for sample units includes sampling plan and sample handling procedures. For a given type of food sampled under NFNAP, numerous brands of processed foods or varieties of raw produce are designated for pick-up at retail outlets or other places of procurement. Each brand or variety may be represented by multiple sample units. Sample units are shipped under proper storage conditions to the Food Analysis Laboratory Control Center (FALCC) at Virginia Polytechnic Institute and State University, Blacksburg, VA. At FALCC, each sample unit receives an identification number; this and other important information including pick-up location and date are recorded.

Composites – Composites are a combination of one or more sample units. In the NDBS, the composite item is used to link all the data contained in the related sample units and subsamples through the Associate function. This provides a mechanism to track a data value at the Compiled level back to the procured sample unit. The

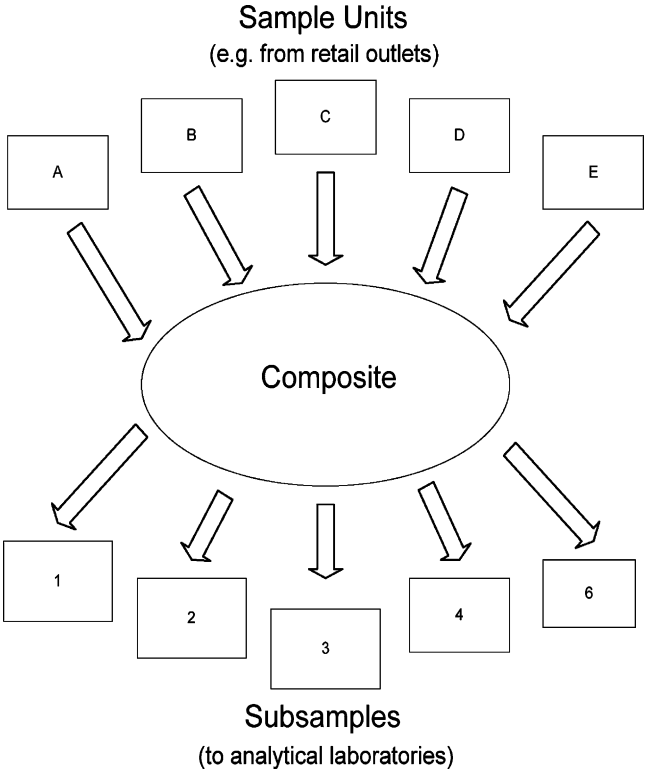

Fig. 4. Data association: sample units, composites and subsamples.

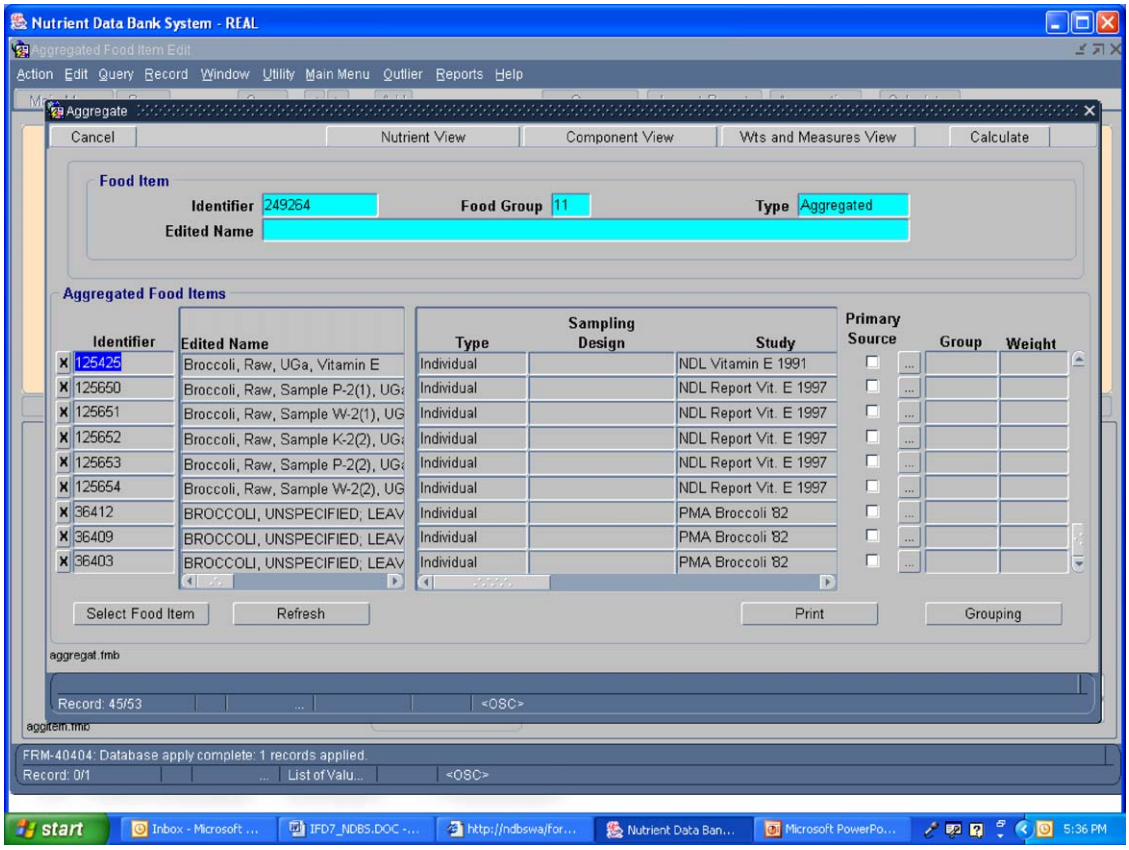

Fig. 5. Sample Aggregation form showing selected initial data.

physical counterpart to this virtual composite in the database is composited and homogenized at FALCC from one or more sample units according to pre-defined work plans and standard protocols.

Subsample – Data for subsamples consist of nutrient values from the analytical laboratory that received the subsample, along with specific analytical methodology information, e.g., method reference and modifications, limits of detection (LOD), and limits of quantification (LOQ). The physical counterparts of the subsamples in the NDBS are a number of aliquots of approximately 25 g each taken from each composited homogeneous mixture and shipped to a variety of laboratories previously certified for the analysis of a particular nutrient or set of nutrients. The aliquots shipped to labs are called subsamples in the databank system and are identified with specific codes or identification numbers.

USDA source codes and the more specific derivation codes are automatically assigned by the system according to algorithms developed during the system design. These algorithms use data stored in Source Category and the Methods detail form for each nutrient on the nutrient tab. Data from the NFNAP program have fully documented methodology fields and will usually be assigned a source code of "1" (analytical) and a derivation code of "A" based on the system finding data in key fields for sample plan, sample handling, and analytical method. Other analytical data, where less information about the method of analysis is available, are assigned different codes. For example, analytical data supplied by a food company that lacks full documentation on methods of analysis, sampling plan, and sample handling are given a source code of 12 and a derivation code of "MA" for Manufacturer's Analytical. There are over 50 derivation codes that can automatically be assigned by the system at various levels from the Initial module through the Compiled module. Approximately 3/4 of the derivation codes describe an AIM_NDBS imputation method. For example, the code "BFSN" indicates that the value was B(ased) on another F(ood) with S(olids) adjustment and N(o) retention factors used. Additional information on source and derivation codes is available in the documentation, which accompanies each release of SR (NDL, 2007a). Files containing all source and derivation codes are part of each release.

2.2. Aggregation module: combining initial data

2.2.1. Basic aggregation

At the Aggregation tier or module, nutrient/food component, weight, and physical component data from a variety of Initial items are combined into a single Aggregation item. In the most simplistic case of a brand name profile supplied by the manufacturer, the Aggregation item may consist of only one Initial item. For generic items or most items analyzed under NFNAP, however, Aggregation brings together data from numerous initial items. Sometimes the Aggregation item may even represent an aggregation of data from a variety of sources such as data from USDA contract analyses, industry trade associations, and the scientific literature. Standard

Table 3
Description of specific tabs available within the Aggregation module.

Tab	Description
Name	Common names for the food item, and information derived from the sampling tab at Initial, such as the regions where the food samples were collected
Weight	Data on common measures entered at Initial are aggregated
Nutrients	Nutrient values of the Aggregation item are displayed
Physical components	Data on components (e.g., bone) entered at Initial are aggregated
LanguaL	LanguaL terms for describing the food item (EuroFIR, 2007)

operating procedures developed during the system design were incorporated into the NDBS to preserve the quality of data, using the source codes assigned at the Initial module. Thus, A, AS (analytical by summation of nutrients/food components) and MA data can be aggregated. However, the system does not allow analytical and calculated data to be aggregated.

Like the form at Initial, the form at Aggregation has three parts. Less information is maintained at the Aggregation level, because data from various sources have been combined and some specificity has been lost; however, the data are still stored within the Initial module and the system maintains links to make them easily accessible. Specific tabs for documenting various types of information within the Aggregation module are shown in Table 3.

At the beginning of the Aggregation process, a query screen is used to select items from Initial to Aggregate, i.e., bring together. Queries are limited to the food group entered on the Aggregation form and are performed using a string search on the edited name field in the Initial module. Other fields in the Initial record can be used in the query to refine the search. For example, a search on "broccoli" would present a list containing any food item with an edited name in food group 11 containing the string "broccoli". This would include many forms of broccoli, such as raw, cooked (by several methods), or frozen. The search can be refined by adding additional terms, e.g., "raw". After the system has identified all the items with the keywords "broccoli" and "raw", the NDL scientists can select which ones to include in the aggregation. The items selected to aggregate are displayed on the form (Fig. 5).

At the Aggregation level, decisions are made on how to group and/or weight the data, combining data from different sources or a single source, or weighting data by market share or production information. The system default is equal weighting.

2.2.2. Specialized statistical procedures used in aggregation

Aggregations are performed using algorithms that are dependent on the type of data being aggregated. Here, descriptive statistics about the data are generated. These include the number of data points that are combined to create the aggregated value, the standard error, the number of different studies represented in the aggregation, the minimum value, the maximum value, the degrees of freedom, and the lower and upper error bounds as appropriate. Error bounds are currently calculated at the 95% confidence level. Choosing the appropriate algorithm allows for the most reliable estimate of the standard error for a specific mean. The specialized statistical procedures developed during system design, along with descriptions, are:

1. Error Bound 1 (EB1) – Algorithm used where there are either two or more study means for each data source or a complete set of observations for all studies from all data sources (e.g., USDA sponsored studies, manufacturers, trade associations, and scientific literature). If two or more studies are not available for a data source, studies from similar data sources having similar weighting factors may be grouped together to create a simulated data source within the NDBS that has at least two studies.
2. Error Bound 2 (EB2) – Algorithm used where only a single observation from at least one data source (e.g., USDA sponsored studies, manufacturers, trade associations, and scientific literature) is available. If each data source has only a single observation, at least two data sources must be used. EB2 should only be used when at least one data source has only a single observation and it is not appropriate to combine some of the data sources to create a simulated data source having at least two observations each.
3. Error Bound 3 (EB3) – Algorithm is used where either all individual values or complete summary statistics for the study (mean, *n*, standard error of the mean) are available for all studies from all data sources (e.g., USDA sponsored studies, manufacturers, trade associations, scientific literature, etc.). Each study must be based on at least two observations (e.g., at least two individual observations per study or a study mean, a standard error and "*n*" for each study). Each observation associated with a study must be based on: (1) an independent simple random sample from the data sources' complete population; and (2) an independent analysis. Analyses should be separated by time and/or laboratory. If this procedure is used when the criteria are not met, the estimated standard error and EB will almost always be too small. Self-weighting data are collected under a sample design (Perry et al., 2003) that ensures the data are self-weighting except for minor deviations and can be treated as a simple random sample from the population of interest (e.g., a particular food sample collected under NFNAP).

The appropriate statistical algorithm described above is selected. A "scenario" screen with online help is available to assist in selecting the correct statistical procedure. Algorithms for imputing values for trace and "not detected" measurements (based on analytical lab-specific LOD or LOQ) are incorporated into the system. A system for documenting statistical procedures (e.g., automatically generated footnotes) used in Aggregation is also included.

In addition to customized aggregation procedures, procedures for testing outliers and comparing sets of data are available. If more sophisticated statistical analysis is required, data can be exported from the database and analyzed using SAS® (SAS, Inc., Cary, NC).

Weight data and physical component data are also aggregated. Any Initial weights having the same number of units, same unit, and same modifier will be aggregated. Thus, all measurements for frozen broccoli that are gram weights for "1 cup, chopped" would be aggregated to give a mean weight for 1 cup of chopped broccoli. A 1-cup weight with a modifier of "florets" would be a separate aggregated mean weight.

2.3. Compiled: finalizing data for dissemination

At the Compiled tier, all the data elements for the food item are finalized. The item is marked to indicate that it is to be included in SR or the dataset used to develop the Food and Nutrient Dataset for Dietary Surveys (FNDDS) (FSRG, 2006). Any missing values are imputed according to procedures described below. The final nutrient profile is selected from aggregation values and imputed values on a Nutrient View screen at Compiled. Final weights and measures and physical component data are selected and dissemination text added as needed. Alternative food product names and codes may also be designated at this level for release. The data form at the Compiled level contains the same three areas (menu bar, header, and information tabs) as the other forms: the specific tabs featured at Compiled are the same as those featured at Initial. However, as most items at Compiled are aggregations of a number of Initial records, some of the fields behind these tabs contain minimal information. For example, since the sample handling occurred on the samples entered into Initial and would be different for different samples, there is no information for this tab at the Compiled level. If a Compiled item is based on a single Initial item, i.e., data from a food company for a brand name item, this information would be carried through the system and populate the appropriate fields at Compiled.

2.3.1. Imputing missing values

2.3.1.1. Imputation methods for commodity items. A wide variety of missing nutrient values can be calculated according to standar-

dized principles (Schakel et al., 1998). This is particularly important for food items that populate the FNDDS, which requires a complete set of values for 64 nutrients/food components and cannot have any missing values. If needed, missing values can be imputed using the recipe or formulation modules within the NDBS, described below. For agricultural commodity items, and less frequently for multi-ingredient items, an imputation method is chosen from a list of imputation procedures, e.g., imputing from a similar food, adjusting for drained solids, or using a mean value for a food category. The NDL scientist then selects the nutrients to impute, the item from which the values are to be imputed, the application of retention factors, and adjustment factors, such as solids basis or fat basis. Appropriate source and derivation codes (described above) are assigned to the imputed value indicating how it was calculated.

Amino acid values are calculated from an amino acid profile. For example, all the amino acid data for multiple forms of broccoli (raw, frozen, and cooked) were aggregated and stored as a profile on the basis of grams of amino acid per gram of nitrogen. At Compiled, this profile is converted to grams of amino acids per 100 g of food through a specialized procedure in the imputing process.

Occasionally, a nutrient value is entered directly, rather than using one of the standard imputation methods integrated into the system. One of the more common instances occurs when the value is assumed to be zero, such as cholesterol in fruit or vegetables. In all cases, an appropriate derivation code must be selected. When a value comes from another food composition database, the source must be documented. If "Other" is chosen, textual documentation of how the value was determined is required.

The item is then compiled and the final nutrient profile appears on the nutrient tab. If any changes are needed in any nutrient values selected, or if the Missing Nutrient Report reveals any missing values for an item intended for use in developing the FNDDS, the NDL scientist can go back to the Compilation screen, use additional imputing methods, or change nutrient value selections.

2.3.1.2. Recipes and formulations. An important method for estimating (i.e., imputing) missing nutrient values in multi-ingredient foods is to use the recipe application, which is part of the databank system. Nutrient values are estimated by calculating them from known ingredient weights; adjustments for moisture and fat losses and gains; and adjustments for nutrient losses due to preparation practices (e.g., heat or exposure to air). An example of recipe use is the preparation of packaged chocolate pudding mix by heating the mix with three cups of milk in a saucepan. The nutrient databank numbers for the mix and milk are selected, and package and cup (volume) measures entered; the system automatically obtains the corresponding weights from data stored in the system. Evaporation losses and nutrient losses due to heat are accounted for by entering the weight loss for moisture and a retention code that triggers the application of expected percent nutrient retentions in heated milk.

The formulation application, which is employed primarily for commercial multi-ingredient foods, uses the recipe application for final calculations. However, it has front-end calculations using the linear programming module of SAS®, allowing the formulation (recipe) for a commercial multi-ingredient food to be estimated. After a formulation "recipe" is generated that meets the constraints entered by the NDL scientist, nutrient values are estimated by the system using the generated ingredient proportions as the recipe.

A team of NDL scientists experienced with ingredient-based estimations, in conjunction with an expert in designing and programming ingredient-based imputation procedures (Westrich et al., 1994), identified new functionalities desirable in a formulation program for the AIM_NDBS system. One of the capabilities identified was the automation of data flow in and out of the formulation program at various stages in the development of a nutrient profile. Therefore, the formulation program was designed to be accessible directly from the main menu as well as from the Compiled module as one of the standard imputing procedures. Ingredient lists that were entered in the Initial module can be displayed in the ingredient tab of the formulation application. This allows the food specialist to easily select SR ingredients that most closely approximate those shown in the food label. As with the recipe application, nutrient losses due to preparation procedures can be approximated by applying appropriate nutrient retention factors (NDL, 2007b). Because the formulation and recipe applications are fully integrated into the AIM_NDBS, recently updated ingredient data can be used in calculations as soon as the ingredient has passed the QC review and been approved for release, by the NDL scientist.

The formulation application can be guided in its formulation estimation by a variety of constraints, most of which are applied on the ingredient tab. They are:

1. Order of ingredient predominance forces the estimated formulation to reflect the label's descending weight order of ingredients.
2. Retention nutrient factors that account for cooking/preparation losses can be applied to ingredients or the product as a whole.
3. Lower and upper bounds can be mandated for a particular ingredient percentage; formulary books and standards of identity (FDA/CFSAN, 2007) help establish bounds.
4. Fat and moisture gains and losses can be applied to individual ingredients or the whole product.
5. Selection of nutrients for best fit matching is done on the target tab.

For the formulation program to be used for a particular food, there must be some known nutrient values. These nutrient values generally come from NFNAP or from analytical or calculated values provided by the manufacturer. The formulation program estimates ingredient percentages that would produce a food with nutrient values similar to a selection of known values. All of the known nutrient values are displayed in the formulation tab called "Target". From these, a limited number of nutrients for matching purposes are selected and targeted. A model error (i.e., estimation of the difference between the known and estimated values) is calculated for each nutrient. However, the total model error represents the summation of the absolute value of the model errors of the targeted nutrients only.

The calculation for an individual nutrient model error is

$$\%\,\text{nutrient model error} = \frac{100\,(\text{estimated value} - \text{known value})}{\text{known value}}$$

This error percent is used to guide revisions to the formulation and qualify the accuracy of the estimates.

2.3.2. Quality control/data validation

A QC report must be run before the item can be approved for release. For example the report alerts the user if the sum of the proximates per 100 g exceeds 100 g or if the sum of the individual sugars exceeds the value for total sugar. A full list of the QC checks is given in Table 4. These checks were developed by NDL scientists based on their experience in compiling food composition data and consider the variability and precision of the various analytical methods. Since a nutrient profile for a Compiled Food Item can be an aggregation of data from many sources, it often happens that summations produce unexpected, or seemingly inconsistent,

Table 4
List of quality control checks performed prior to release of each version of the USDA National Nutrient Database for Standard Reference.

Nutrient class	Quality control check
Proximates	Sum of values for moisture, protein, fat, carbohydrates, ash, and alcohol should be between 99.8 and 100.2 for plant products For meat products, the acceptable range is 97–103
Carbohydrates	Sum of values for total dietary fiber, total sugar and starch should not exceed value for total carbohydrates Values for any one of the carbohydrate fractions should not exceed the value for total carbohydrates Values for the sum of individual sugars (sucrose, glucose, fructose, lactose, maltose, and galactose) should not exceed the values for total sugars
Calories	Kilojoules should equal kilocalories × 4.184
Fat and fatty acids	Sum of values for total saturated fatty acids, total monounsaturated fatty acids, and total polyunsaturated fatty acids should not exceed values for total fat Sum of individual fatty acids should not exceed values for total fat Sum of individual *trans* fatty acids should not exceed the value for *total* trans fatty acids
Minerals	Sum of minerals does not exceed the value for ash
Folate	Sum of food folate and folic acid should not exceed value for total folate Value for food folate plus 1.7 × the folic acid value should not exceed the value for dietary folate equivalents
Vitamin B12	Value for added vitamin B12 should not exceed the value for total vitamin B12 Should not have a value for added vitamin B12 if do not have a value for total vitamin B12
Vitamin E	Value for added vitamin E should not exceed the value for total vitamin E Should not have a value for added vitamin E if do not have a value for total vitamin E
Vitamin A	Sum of individual carotenoids (α-carotene, β-carotene, β-cryptoxanthin) times appropriate factor should be equal or less than the value for total vitamin A in IU or RAE Vitamin A in RAE should be less than the value for vitamin A in IU Value for retinol in μg should be less than value for vitamin A in IU Value for retinol in μg should be less than or equal to the value for vitamin A in RAE
Phytosterols	Sum of stigmasterol, campesterol and β-sitosterol should not exceed the value for total phytosterols
Other	If the refuse value is zero, there should be no description in the refuse description field Each food item should have a refuse value Each food item should have at least one household measure If the source code = 7, which indicates an assumed zero, the nutrient value should be zero

results. For example, carbohydrate determined by summation of analytical values for individual sugars, starch, and other carbohydrate fractions may not equal carbohydrate by difference, but should be within 0.5 g per 100 g of food.

2.3.3. Approvals

After running the QC checks and making any indicated changes, the approval box is checked. The final approval step is a technical review (TR approval). Prior to release, the data are sent to experts for review. In the case of brand name items, this is the food manufacturer. For other foods, external experts familiar with the food and its nutrients are recruited to review the data. Clicking the TR approval box generates a comment form in which information such as reviewer names and comments can be entered. For food items in the FNDDS, a change code form for the documentation of any nutrient value changes since the last release is displayed. If any of the 64 mandatory survey nutrients/food components are missing, an alert is issued.

3. Dissemination and products

A number of utilities are available to disseminate the four principal files (food description, nutrient data, weight, and footnotes) as well as the six support files (food group descriptions, nutrient definitions, source codes, data derivation descriptions, sources of data, and sources of data links) that comprise each release of SR. An abbreviated file, which contains all the food items but fewer nutrients, is also disseminated. These files are exported from the NDBS as ASCII delimited files and are made available on NDL's Web site (www.ars.usda.gov/nutrientdata). For the convenience of the user, NDL imports these files into a Microsoft® Access database. The abbreviated file is available as a Microsoft® Excel spreadsheet. Page images comparable to the printed Agriculture Handbook No. 8 (USDA, 1976–1992) are generated from this utility. Data files and reports for nutrient retention factors and food yields are also disseminated at this time, if updates are available.

The first SR to be disseminated from AIM_NDBS was SR14, released in February 2002. Owing to the new capabilities of the databank system, this release had added fields designed to improve descriptive information for food items and provided expanded statistical information about nutrient values. Population of these new fields is an ongoing process as older data are replaced by current data from NFNAP and other sources.

4. Conclusion

The USDA National Nutrient Databank System provides a powerful collection of tools for managing the diverse data collections of the USDA on the composition of foods. The food item table contains over 230,000 records, while the nutrient value table contains over 3.3 million records. Records for Initial (including sample units, subsamples and composites), Aggregation, and Compiled modules are contained in the same database tables, so these figures represent the total number of records for all levels of the system. This system supports the annual releases of SR, which are widely used by other government agencies developing nutrition policy, health researchers and food companies conducting research, and consumers interested in what is in the food they eat. It also provides data issued to create the FNDDS used in the What We Eat in America: NHANES (NCHS, 2006).

Some of the major accomplishments in the development of the new databank system have been:

1. integration of stand-alone functions into one comprehensive database system;
2. flexibility to handle entry of data from a variety of sources, including USDA sponsored contracts, other government agencies, the food industry, and the scientific literature;
3. mechanisms to validate data;
4. mechanisms to perform a variety of different aggregations using specialized statistical algorithms, developed for the system;
5. enhancement of recipe and formulation calculations;
6. improved production of reports for internal use as well as for dissemination to a variety of users.

Since the current database system's inception in 1997, NDL has made numerous enhancements to the system to improve its functionality. These enhancements allow NDL to better meet the requirements of its mission to develop authoritative food composition databases and state-of-the-art methods to acquire, evaluate, compile, and disseminate composition data on foods available in the United States.

Acknowledgements

The authors wish to acknowledge the contributions of the Nutrient Data Laboratory staff in the development of the Nutrient Databank System: Jacob Exler, Susan Gebhardt, Juliette Howe, Gwen Holcomb, Kris Patterson, Bethany Showell, Melissa Nickle, and Robin Thomas. Rena Cutrufelli, in particular, is acknowledged for project coordination and serving as the liaison with contractors. Joanne Holden, Research Leader, is recognized for overall support and leadership.

References

Atwater, W.O., Woods, C.D., 1892. Investigations upon the Chemistry and Economy of Foods. Report Connecticut (Storrs) Agric. Expt. Sta. for 1891.

EuroFIR Consortium, 2007. LanguaL. Retrieved 2007-12-20 from: www.langual.org.

Food Surveys Research Group (FSRG), 2006. Agricultural Research Service, U.S. Department of Agriculture. USDA Food and Nutrient Database for Dietary Studies, 2.0, Beltsville, MD. Retrieved 2006-9-26 from: http://www.ars.usda.gov/Services/docs.htm?docid=12089.

Haytowitz, D.B., Pehrsson, P.R., Holden, J.M., 2007. The National Food and Nutrient Analysis Program: a decade of progress. J. Food Comp. Anal. 21 (Supp. 1), S94–S102.

Holden, J.M., Bhagwat, S.A., Patterson, K.Y., 2002. Development of a Multi-Nutrient Data Quality Evaluation System. J. Food Comp. Anal. 15 (4), 339–348.

Holden, J.M., Bhagwat, S.A., Haytowitz, D., Gebhardt, S., Dwyer, J., Peterson, J., Beecher, G.R., Eldridge, A.L., 2005. Development of a database of critically evaluated flavonoid data: application of USDA's Data Quality Evaluation System. J. Food Comp. Anal. 18, 829–844.

Marcoe, K.K., Haytowitz, D.B., 1993. Estimating nutrient values of mixed dishes from label information. Food Technol. 47 (4), 69–75.

National Center for Health Statistics (NCHS), Center for Disease Control and Prevention (CDC), Department of Health and Human Services (DHHS), 2006. National Health and Nutrition Examination Survey 2003–2004 Data Files. Retrieved 2007-07-02 from: http://www.cdc.gov/nchs/about/major/nhanes/nhanes03-04.htm.

Nutrient Data Laboratory (NDL), Agricultural Research Service (ARS), U.S. Department of Agriculture (USDA), 2007a. National Nutrient Database for Standard Reference, Release 20. Retrieved 2007-09-26 from: http://www.ars.usda.gov/Services/docs.htm?docid=8964.

Nutrient Data Laboratory (NDL), Agricultural Research Service (ARS), U.S. Department of Agriculture (USDA), 2007b. USDA Table of Nutrient Retention Factors, Release 6. NDL Web site: Retrieved 2007-12-20 from: http://www.ars.usda.gov/Main/docs.htm?docid=9448.

Perry, C.R, Pehrsson P.R., Holden J., 2003. A Revised Sampling Plan for Obtaining Food Products for Nutrient Analysis for the USDA National Nutrient Database. 2003 Proceedings of the American Statistical Association (CD-ROM). Retrieved 2007-07-02 from: http://www.amstat.org/ASAStore/BOoks_Proceedings_CDs_C4.cfm.

Schakel, S.F., Buzzard, I.M., Gebhardt, S.E., 1998. Procedures for estimating nutrient values for food composition databases. J. Food Comp. Anal. 10, 102–114.

U.S. Department of Agriculture. Composition of Food: Raw, Processed, Prepared Agric. Handbook No. 8: AH 8 1, Dairy and Egg Products, 1976; AH 8 2, Spices and Herbs, 1977; AH 8 3, Baby Foods, 1978; AH 8 4, Fats and Oils; 1979; AH 8 5, Poultry Products, 1979; AH 8 6, Soups, Sauces, and Gravies, 1980; AH 8 7; Sausages and Luncheon Meats, 1980; AH 8 8, Breakfast Cereals, 1982; AH 8 9, Fruits and Fruit Juices, 1982; AH 8 10, Pork and Pork Products, 1983; AH 8 11, Vegetable and Vegetable Products, 1984; AH 8 12, Nut and Seed Products, 1984; AH 8 13, Beef Products, 1986; AH 8 14, Beverages, 1986; AH 8 15, Finfish and Shellfish Products, 1987; AH 8 16, Legumes and Legume Products, 1986; AH 8 17, Lamb, Veal, and Game Products, 1989; AH 8 20, Cereal Grains and Pasta, 1989; AH 8 21, Fast Foods, 1988; 1989 Supplement, 1990; 1990 Supplement, 1991; 1991 Supplement, 1992.

U.S. Food and Drug Administration/Center for Food Safety and Applied Nutrition (FDA/CFSAN), 2007. What are the requirements regarding food standards? Retrieved 2007-12-20 from: http://www.cfsan.fda.gov/(dms/qa-ind2d.html.

Westrich, B.J., Buzzard, I.M., Gatewood, L.C., McGovern, P.G., 1994. Accuracy and efficiency of estimating nutrient values in commercial food products using mathematical optimization. J. Food Comp. Anal. 7, 223–239.