USDA's FoodData Central: what is it and why is it needed today?

Naomi K Fukagawa, Kyle McKillop, Pamela R Pehrsson, Alanna Moshfegh, James Harnly, and John Finley

USDA Agricultural Research Service, Beltsville Human Nutrition Research Center, Beltsville, MD, USA

ABSTRACT

FoodData Central (FDC) is the center of the USDA-based food-composition information web. It is an integrated data system that presently provides—in 1 place—5 distinct types of data containing information on food and nutrient profiles. Each data type has a unique purpose. Two of the data types—Foundation Foods (FF) and Experimental Foods (EF)—represent "a bridge to the future" in food and nutrient composition. They provide data and metadata that have never previously been available from a database. The other 3 data types are well established and familiar to many users: Standard Reference (SR) Legacy, Food and Nutrient Database for Dietary Studies (FNDDS), and Global Branded Foods Products Database (GBFPD). After >100 y of maintaining food-composition data within the USDA, it was clear that change was needed to respond to the rapid increase in the number and variety of foods in the food supply, evolution of analytical approaches, and new agricultural practices and products. FDC is USDA's answer to the challenge of providing reliable, web-based, transparent, and easily accessible information about the nutrients and other components of foods to meet the increasingly diverse needs of many audiences, including public health professionals, agricultural and environmental researchers, policy makers, nutrition professionals, health care providers, product developers, and the public at large. *Am J Clin Nutr* 2022;115:619–624.

Keywords: food composition, database, nutrients, USDA Agricultural Research Service, food supply, food analysis, food products, diet, FoodData Central

Introduction

The USDA was the first federal agency to conduct human nutrition research. This mission started in 1894 when the *Yearbook of the USDA* included Wilbur O Atwater's work in human nutrition and demonstrated its importance to public health, medical sciences, and the military. In 1895, Atwater published a USDA bulletin that summarized the available data on food composition, food digestibility, known nutritional needs, and protocols for these investigations (1). For >100 y since then, USDA's Agricultural Research Service (ARS) has continued to analyze foods and determine the consumption and dietary patterns of Americans. This work has kept pace with the evolution in analytical approaches, technology, and agricultural practices that have occurred since Atwater first published his findings. However, the food supply and the scientific understanding of interrelations between dietary intakes, food components, and health have grown significantly over the years. At the same time, changes in food production, composition, and availability have contributed greatly to the complexity of the relation between food consumption and public health. The massive increase in the number of prepared and proprietary food products available to consumers as well as rapid and frequent changes in formulations or recipes have made it impossible to analytically determine the composition of all, or even the most important, foods in the food system. This has highlighted the importance of understanding the composition of key ingredients used to make these foods.

There are many challenges related to understanding the interaction of agricultural practices, food composition, diet, and health, including how domestic and international food composition is determined and reported. These challenges, as well as the explosive growth of the food supply, rapid product changes, evolution and access of computer technology and their applications, and growing consumer interest in diet and health, have necessitated re-examination of the data provided to the public by the USDA ARS. This re-examination has also created challenges for individuals who want healthy, accessible, and affordable diets for themselves and their families. Consequently, there is urgency for USDA's food-composition data resources to evolve to meet the increasingly diverse needs of many audiences, including public health professionals, agricultural and environmental researchers, policy makers, nutrition professionals, health care providers, product developers, and the public at large. These

All authors are employed by the USDA Agricultural Research Service and have not received external support in the preparation of this manuscript.

Address correspondence to NKF (e-mail: naomi.fukagawa@usda.gov).

Abbreviations used: AgCROS, Agricultural Collaborative Research Outcomes System; ARS, Agricultural Research Service; FDC, FoodData Central; FF, Foundation Foods; FNDDS, Food and Nutrient Database for Dietary Studies; GBFPD, Global Branded Food Products Database; SR, Standard Reference.

Received August 8, 2021. Accepted for publication November 25, 2021.

First published online December 10, 2021; doi: https://doi.org/10.1093/ajcn/nqab397.

data resources must be publicly available and provide web-based, transparent, and easily accessible information about the composition of foods. In addition, these resources must provide information regarding the factors that may affect compositional variability, including genetics, environment, management, production, and processing systems and how they are linked to indices or biomarkers of health. Finally, the resource needs to be flexible as definitions evolve and data are accessed and used in multiple, diverse arenas.

The need for a dramatic change in the USDA National Nutrient Database for Standard Reference (SR) was recognized by the USDA Office of National Programs and the Beltsville Human Nutrition Center in 2012. It was understood that the existing approach to assembling the SR was no longer viable given the demands of the US food supply and the limits of USDA's resources, even with the generous support of other federal agencies. The first step was an infusion of new funds and the establishment of a Database Modernization Committee. The second step was the establishment in September 2016 of a private–public partnership, a collaboration between USDA, Institute for the Advancement of Food and Nutrition Sciences (formerly ILSI North America), GS1 US, 1WorldSync, Label Insight, and the University of Maryland. The partnership resulted in the creation of the Global Branded Food Products Database (GBFPD) based on commercial label data.

The advent of GBFPD prompted a change in the USDA's internal approach to data collection, analysis, and transparent presentation. Updates to SR were halted and the last version was labeled SR Legacy and released in 2018. The focus of USDA's internal data acquisition and analysis shifted to commodities and single-ingredient foods, or the foundation foods that are the building blocks of the complex foods in the GBFPD. The new Foundation Foods (FF) data type concentrated on the variability of the raw foods in the US food supply as opposed to the previous market-basket survey approach that was based on market share. The focus on variability produces an emphasis on the metadata associated with foods, i.e., genetics, environment, management, and processing. The emphasis is on the analysis of individual samples of foods rather than composites. The user can easily access the variety, location, time of acquisition, and other critical metadata associated with each food. The result is a focus on single-ingredient foods and a dramatic increase in the amount of data for each food.

The evolution of GBFPD and FF also identified 2 additional areas of expansion and consolidation of data accessibility. Food-composition data within the context of an experimental design or derived from new analytical methodology research do not typically appear in public databases but would be beneficial to researchers if more broadly accessible. That need identified data for Experimental Foods (EF). Another data type developed by Beltsville Human Nutrition Research Center's (BHNRC's) Food Surveys Research Group and specific to national nutrition monitoring is the Food and Nutrient Database for Dietary Studies (FNDDS) that has been released every 2 y in concert with the 2-y release of What We Eat in America as part of the NHANES. Including FNDDS as a data type supported the goal of centralized availability. Together, these 5 data types comprise USDA's FoodData Central (FDC).

FoodData Central, USDA's Response to Food- and Nutrient-Composition Data Needs

FDC, the center of a USDA-based food-composition information web, is an integrated data system that presently provides—in 1 place—5 distinct types of data containing information on food and nutrient profiles. Each data type has a unique purpose. Two of the data types—FF and EF—represent "a bridge to the future" in food and nutrient composition. They provide data and metadata that have never previously been available, as follows:

- FF provides information on single-ingredient foods and basic food ingredients used to compose more complex food items. FF stresses analytical values for components of individual food samples without compositing of multiple samples as previously done. The analyses include defined nutrients as well as other nutritionally important compounds. This approach provides information on the variability of the food supply and includes extensive underlying metadata, such as the number of samples, sampling location, date of collection, analytical approaches used, and if appropriate, agricultural information such as genotype, growing location, and production practices. In contrast to FF, SR Legacy presents average values from selected foods in the marketplace and derived from analyses of composites, calculations, and/or from the literature. Recognition that the food system is more complex and fluid necessitated the change in approach to selecting and analyzing foods to be included in FF. The enhanced depth and transparency of FF data can provide valuable insights into the many factors that influence variability in nutrient and food component profiles. Unlike the large number of other items in FDC, FF emphasizes single-ingredient foods and agricultural commodities that constitute the building blocks of our complex system of foods. The goal of FF is to provide information on the foods used to make the more complex foods that are consumed. To achieve this, the number of ingredients that are used to make foods as consumed will increase over time; this includes single agricultural commodities such as corn, potatoes, eggs, beef, and poultry; commonly used basic food ingredients that have been processed such as flours, some cheeses, and cooking oils; and some food additives such as gums and soy lecithin. Less emphasis will be placed on mixed dishes or prepared foods. The inclusion of metadata related to FF will facilitate researchers' ability to link food-composition data with other important USDA datasets [e.g., Nutrient Uptake and Outcomes Network, the Agricultural Collaborative Research Outcomes System (AgCROS), and Economic Research Service]. FF represents the future of USDA food information because of access to metadata and links to agricultural and production information available across USDA and other federal agencies. One of the goals is to develop systems to include information from nongovernmental sources.
- EF contains foods produced, acquired, or studied under unique conditions, such as alternative management systems, experimental genotypes, or research/analytical protocols. Emerging data show that genetics, the agricultural system, and the environment can have a significant impact on

the composition of foods. Data for EF do not generally fit well into any of the other data types. The foods in this data type may not be commercially available to the public, but the data may expand information about a specific food or the research contributes nutritional insights applicable to humans. Often, these data are published in other research disciplines (e.g., chemistry, genetics, plant physiology, animal management, climate) and may not be easily accessible to those focused on human nutrition and food composition. EF are for research purposes and may not be appropriate as a reference for the consumer or for diet planning. EF data may also be available through links to relevant agricultural research data sources, such as the AgCROS. The data in EF may include (or link to) variables such as genetics, environmental inputs and outputs, supply chains, economic considerations, and nutrition research. These data will allow users to examine a range of factors used that may affect the profiles of food components, including nutrients, and resulting dietary intakes as well as the sustainability of agricultural and dietary food systems. Knowing the variation in single foods underscores the challenges of understanding exactly what and how much is being consumed and their biological effects. It is clear that foods on a plate interact, emphasizing the importance of new approaches to elucidate dietary patterns and identify synergistic effects that are likely to promote wellness, especially as we move towards precision nutrition and precision medicine.

The other 3 data types in FDC are well established and familiar to many users:

- SR has been the primary food-composition data type in the United States for decades. SR provides a comprehensive list of values for nutrients and food components that are derived from analysis, calculations, and/or the literature. Analytical nutrient profiles were determined for foods in the marketplace and selected based on the market share and thus an estimate of average exposure across the United States. This data type has provided the values for most other public and private food-composition databases and has supported a wide range of public policy initiatives, research studies, diet planning, and education activities. SR Legacy, released in April 2018, is the final release of this data type. Although access to SR will be maintained, the values will become less relevant as the food supply evolves over time. Data in the GBFPD and FF are expected to provide more timely information on the food supply.
- FNDDS provides nutrient values for the foods and beverages reported in What We Eat in America, the dietary intake component of the NHANES (2). FNDDS data facilitate analyses of dietary intakes reported in NHANES as well as many other dietary research studies. The nutrient profiles of a majority of foods and beverages in FNDDS 2017–2018 were generated using a recipe calculation process utilizing 2 or more ingredient codes from FF and SR Legacy data in FDC. FNDDS includes documentation of the FDC source of each nutrient value for ingredients used in FNDDS and is publicly available. While SR Legacy accounts for most of the ingredient nutrient profiles, it is anticipated that FF use in FNDDS will increase as FF data expand. FNDDS includes documentation of the FDC data source for each nutrient value for all foods in this data type. FNDDS and related database resources developed to enhance dietary data analysis continue to be accessible on the Food Surveys Research Group website in addition to FNDDS being available in FDC.
- GBFPD includes data from a private–public partnership whose goal is to enhance the open sharing of nutrient data that appear on branded and private label foods and are voluntarily provided by the food industry. Members of this partnership are as follows: *1*) USDA's ARS, *2*) Institute for the Advancement of Food and Nutrition Science, *3*) GS1, *4*) 1WorldSync, *5*) Label Insight, and *6*) the University of Maryland, Joint Institute for Food Safety and Applied Nutrition. Information in GBFPD is received from food industry data providers and displayed after conversion to 100-g measures to align the presentation with the rest of FDC. GBFPD contains information on more than several hundred thousand foods with monthly updates. The dataset focuses on collecting nutrients as required on the label specified by FDA in the Federal Register. The capacity exists for data providers to supply additional, more extensive data about the food beyond what is required in the label. Inclusion of this data type in FDC has enabled USDA to focus on analytical food-composition research with expanded representation of food components, including nutrients, and associated metadata for a food.

The Future of FDC

New FF and EF data will be added to FDC semi-annually and will include data from multiple sources, with an emphasis on single-ingredient foods and those that have been produced under experimental conditions. SR Legacy data will remain static. The data in EF include (or link to) variables such as genetics, environmental inputs and outputs, supply chains, and economic considerations. These data will allow users to examine a range of factors, such as cultivar, geography, and agricultural or experimental methods used, that may affect the nutritional profiles of foods and resulting dietary intakes as well as the sustainability of agricultural and dietary food systems. In addition to semi-annual additions of new data, enhanced features for data viewing and extraction will be incorporated periodically.

The differences between the data available from SR and FF as well as the enhanced usability of FF data are illustrated in **Figures 1** and **2**. Figure 1 shows data for total sugars in bananas from SR and recently published data from FF. In addition to access to values from individual samples and additional metadata (e.g., source) in FF, the figure shows how variable the total sugar content can be and dependent on source and ripeness.

The impact of genetics and environment on nutritional composition of foods is illustrated in Figure 2 by data for "pinto beans, dry" located in FF. This type of visual representation from the data in FF shows the distribution of iron content among different cultivars, with the growing location being the dominant contributor to variability. The frequency plots show 60 genotypes harvested from 2 locations (Colorado and Puerto Rico). Pinto beans are considered an important source of protein, fiber, vitamins, and minerals. The data reported in FDC show that

	SR	FF
Mean	12.2	15.8
Min	7.5	12.7
MAX	16.2	20.1
S1	--	18.1
S2	--	14.8
S3	--	16.3
S4	--	20.1
S5	--	13.9
S6	--	17.0
S7	--	12.7
S8	--	15.9
S9	--	15.7
S10	--	15.8
S11	--	15.3
S12	--	14.3

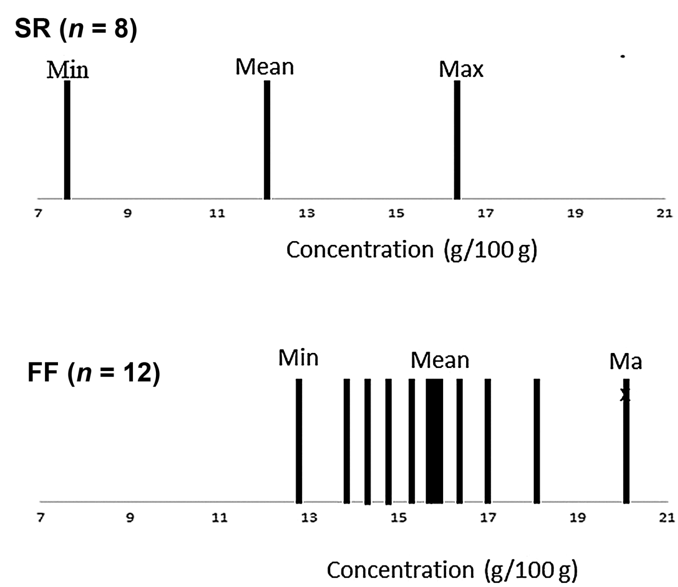

FIGURE 1 Data for total sugars in bananas from SR and FF are shown in the table on the left. Frequency plots of the data are shown on the right, SR (top) and FF (bottom). The SR database shows only the minimum (min), maximum (max) and the mean for the 8 samples analyzed. Data for individual samples are not available. For FF, data are available for the 12 samples analyzed and all 12 are shown on the frequency plot. Frequency resolution is 0.1 g/100 g. Plots for FF samples at 15.7, 15.8, and 15.9 appear as a triple thick line. FF, Foundation Foods; SR, Standard Reference.

iron concentrations varied from 3.4 to 7.7 mg/100 g (∼226% change), calcium from 93 to 246 mg/100 g (∼264% change), and total lipids from 0.83 to 1.74 g/100 g (∼210% change). Figure 2 shows that the iron content of the pinto beans appears to be primarily impacted by location. This would mean that the source of the beans on your plate could affect your intake of iron. If your beans were sourced from the lower end of the range, you would be receiving ∼37% less iron per 100 g than if you assumed that the content was the reported average of 5.4 mg/100 g; if sourced from the upper end of the range, you could consume ∼43% more than the 5.4-mg/100-g average reported in FDC. This may not be significant if you are healthy and have access to a varied diet, but it may be an issue if you are iron-deficient, have a parasitic infection, or have poor access to sufficient, healthy food. Conversely, one must also consider unintended adverse consequences of excessive intake of iron depending on the other foods in your diet. FDC's FF reports data from analyses of individual samples and encourages users to consider variation due to a number of factors as they decide which value to use in the context of their application of food-composition data. The

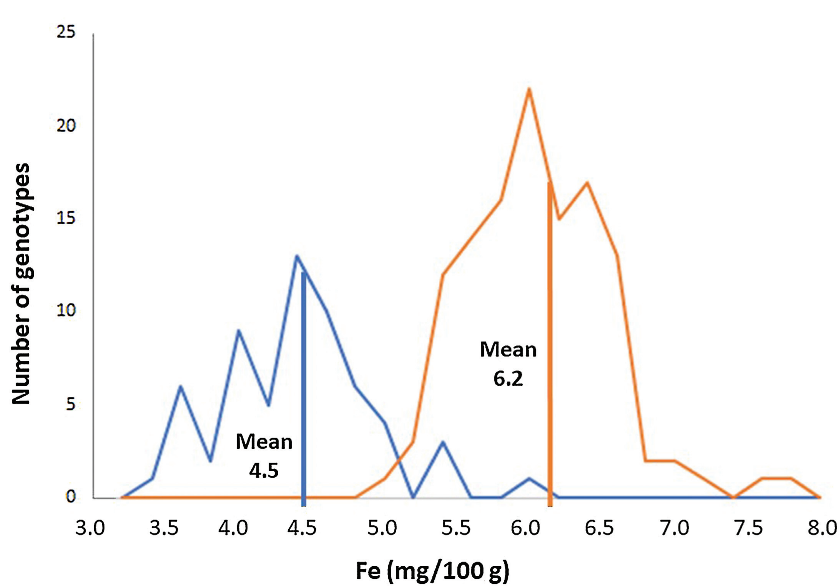

FIGURE 2 Frequency plots for iron (Fe) in pinto beans grown in Colorado and Puerto Rico. Plots show the same 30 genotypes analyzed in duplicate for Colorado (60 data points), shown in blue, and in quadruplicate for Puerto Rico (120 data points), shown in orange. Frequency resolution is 2 g/100 g.

format of data entry and presentation in FF allows researchers to better understand many of the parameters of variability and not rely on a single value. As genetics, environment and agricultural, management, and processing practices evolve and are better understood, scientists can better plan for the future global food supply.

Keeping Pace with Changes in Sources of Evidence and Analytic Approaches

Traditional sources of data for food profiles have included the scientific literature, food industry, laboratory analyses of select foods, calculations from ingredient lists or recipes, or estimations based on similar foods. There has been increasing recognition that profiles of food components are impacted by the rapid pace of change in the food supply; the inherent variability in data resulting from factors such as cultivar and breed, maturity and age, and color; environmental factors, such as climate and agricultural management practices; and food supply issues, such as processing, storage, transportation, and preparation. Hence, USDA acknowledges that its food data capability must be flexible, responsive, and accurate. The need for such a system has been exemplified by unexpected system shocks such as the coronavirus disease 2019 (COVID-19) pandemic. FDC is a key step forward in this process of providing a flexible and responsive data system that allows diverse audiences to access a variety of data types, obtain data on variability in nutrient and food component values, and connect between different databases (each designed with a different purpose in mind) through an emerging ontology that provides linked terminology for food names, descriptors, chemical compounds, new food components identified, and other relevant factors.

FDC is establishing new ways to incorporate data. The old paradigm of careful targeted analysis of a complete set of nutrients does not work in a time when analysis of 1 food can exceed $50,000 and when hundreds to thousands of nonnutritive components with bioactivity are known. Moreover, the food supply is in constant flux: new products are added, old products are removed, existing products are reformulated, new cultivars or breeds are introduced, and new foods are developed. Hence, nontargeted, multicomponent analyses will become more common. Any form of food-profile data must be understood to simply be a "snapshot in time," as no established "gold standard" exists for these data.

The data types currently in FDC provide a greatly enhanced spectrum of food-profile information, but each data type has its own challenges, which opens the door to considerations about other data types that could be added in the future. The GBFPD contains food label information provided by the public–private partnership and is continuously updated as the products change in the marketplace. However, the compositions of most foods in this dataset are limited to label information and the data are presented as they are submitted by their data providers. FF provides greater breadth with respect to the number of analyzed components and sampling to characterize variability, but analytical costs limit the number of foods that can be analyzed. EF provides data on the potential variability introduced by agricultural and processing technology, and analytical techniques. Experimental foods are not necessarily the same as commercially available foods but may help to spawn innovation in food production, distribution, and consumption.

Flexibility with regard to the data types must also be responsive to the evolving uses of the data. In the past, food-composition data were used primarily to link nutrient intakes to health outcomes, especially through What We Eat in America, the food intake component of the NHANES (2). Food-composition data were also used to support epidemiologic surveys, to draw conclusions about the relation of consumption of particular nutrients and foods with health outcomes, and to promulgate policy to encourage consumption of those foods and nutrients associated with positive health outcomes. However, many of the questions being posed today are different. Consumption of individual foods may be of less importance than dietary patterns. This was confronted by the 2020 Dietary Guidelines Advisory Committee (3). Many reports, such as the Lancet EAT Commission (4), have advocated radical changes in diets for both human and environmental health; however, adequate modeling of potential dietary-induced changes requires diverse and integrated data sources with associated variability.

Interest in the food system as a whole is increasing and questions are being asked about sustainability and the ability of current food systems to feed a world of a projected 9 to 10 billion people (5). Such questions can be answered only by more complex and integrated models and analyses. FDC is building links, metadata, and an ontology to facilitate such analyses.

The USDA is responsible for providing the best data available and data that are accessible and usable by a wide audience. It will be important that users recognize that, depending on the dataset used, data may vary and be of uneven quality, reliability, and not necessarily representative of the current food supply. The US food supply has become global so the present emphasis on nationally representative data will need to be expanded to be globally representative. Researchers will need to develop new ways to use the data to meet the needs of individuals and to contribute to population and clinical studies where specific intakes are correlated with biological measures that are themselves highly variable. The worldwide food system is undergoing an enormous transformation and will probably be unrecognizable in another 30 y. Contemporary issues facing the food system are not confined to discrete disciplines. For example, much has been written about the need to change the food system to provide diets that are healthy but with less environmental impact. Addressing such questions requires an integrated array of diverse types and forms of data. Successfully addressing questions will be intimately tied to reducing barriers between "silos" of data and establishing incentives for academic and private sectors to accept and implement the concept of open data (6). FDC is designed to have the flexibility to accommodate and evolve with such changes.

The authors' responsibilities were as follows—all authors: contributed to the conception, the draft, revised and final version of the manuscript and read and approved the final manuscript. The authors report no conflicts of interest.

Data Availability

All data that appear in FDC are publically available through the FDC website and application.

References

1. Dupont JL, Beecher GR, editors. History of human nutrition research in the U.S. Department of Agriculture, Agricultural Research Service: people, events and accomplishments [Internet]. ARS-177. Washington (DC): USDA, Agricultural Research Service; 2017. Available from: https://www.ars.usda.gov/ARSUserFiles/oc/np/HistoryofHumanNutritionResearch/HistoryofHumanNutritionResearch.pdf (accessed 3 May 2019).
2. Department of Agriculture, Agricultural Research Service. 2018. USDA Food and Nutrient Database for Dietary Studies 2017-2018 [Internet]. Available from: https://www.ars.usda.gov/northeast-area/beltsville-md-bhnrc/beltsville-human-nutrition-research-center/food-surveys-research-group/docs/fndds-download-databases/.
3. US Department of Agriculture; US Department of Health and Human Services. Dietary guidelines for Americans, 2020-2025. 9th ed. December 2020. [Internet]. Available from: DietaryGuidelines.gov
4. Willett W, Rockström J, Loken B, Springmann M, Lang T, Vermeulen S, Garnett T, Tilman D, DeClerck F, Wood A, et al. Food in the Anthropocene: the EAT-Lancet Commission on healthy diets from sustainable food systems. Lancet North Am Ed 2019;393(10170):447–92. doi: 10.1016/S0140-6736(18)31788-4.
5. United Nations, Department of Economic and Social Affairs, Population Division. World population prospects. New York, New York: United Nations; 2019. Data Booklet (ST/ESA/SER.A/424).
6. Council for Agricultural Science and Technology (CAST). 2019. Enabling open-source data networks in public agricultural research. CAST Commentary QTA2019-1. CAST, Ames, Iowa.