Phoebe Brand

LIS 545 B Data Curation I: Fundamentals

Due: Feb 23, 2024

**Final Term Project: Final Report**

**Data and Metadata Profile**

***Data***

I chose the October 2023 Keydata Report for Nutrition Assistance Programs, [(*October 2023 Keydata Report | Food and Nutrition Service*, 2024)](https://www.zotero.org/google-docs/?rzl0dR) which includes data on the Supplemental Nutrition Assistance Program (SNAP), the Special Supplemental Nutrition Program for Women, Infants, and Children (WIC), child nutrition programs (National School Lunch, School Breakfast, Child and Adult Care, Summer Food Service and Special Milk), and food distribution programs (Schools, Emergency Food Assistance, Indian Reservations, Commodity Supplemental, Nutrition for the Elderly, and Charitable Institutions).

The dataset is published by the Food and Nutrition Service of the U.S. Department of Agriculture (USDA). The data comes from these assistance programs on a state level, describing variables like cost, participation, commodity schools, meals, total costs, entitlement, and center for each program. The title page of the Excel document claims it is “generated from National Data Bank Version 8.2”. I had difficulty finding an explanation for this. Still, I did find a document on calculating SNAP Indexes from 2013 that says: “The monthly counts are extracted from the Food and Nutrition Service National Data Bank”. [(*Calculating the Supplemental Nutrition Assistance Program (SNAP) Program Access Index: A Step-by-Step Guide for 2013*, 2015)](https://www.zotero.org/google-docs/?vIttpi) Each program sheet also lists it as derived from the National Data Bank version 8.2.

There is a wide range of potential stakeholders for this data. The creators are the Food and Nutrition Service of USDA. The data is managed by the FNS Office of Policy Support (OPS) which “leads the development and execution of FNS’s study and evaluation agenda”.[(*Research & Analysis | Food and Nutrition Service*, n.d.)](https://www.zotero.org/google-docs/?6TFbzA) In terms of funding, the FNS funds numerous projects and grants[(*Grants | Food and Nutrition Service*, n.d.)](https://www.zotero.org/google-docs/?4CpjMO) that may utilize this data, and researchers for these projects include those in topics like Assessing/Improving Operations, Food Security, Nutrition Education, Participation Rates and Characteristics, and Payment Accuracy and Program Integrity. [(*Research & Analysis | Food and Nutrition Service*, n.d.)](https://www.zotero.org/google-docs/?tC0IFf) Program partners, like other member divisions of the USDA, may utilize this data. In the policy realm, government officials and programs can utilize these reports to back proposals, as cited in the Reports to Congress section of the USDA Data and Research site. [(*Data & Research | Food and Nutrition Service*, n.d.)](https://www.zotero.org/google-docs/?mBncFL) Other users may include any members of the public, as this is publicly accessible data.

The most recent month’s data (at this time October) Keydata Report is made available as a PDF and an xlsx (listed as a slsx perhaps mistakenly on the site) document. These two formats contain the same data, with 20 sheets or pages. There is another report, the Performance Report, which is only one sheet and a summarized form of the Keydata. Most users should be able to open an xlsx file, and PDFs are even easier.

This data is open access, with no restrictions listed. There is also a page on implementing open-access resources by the FNS. (Food and Nutrition Service Implementation Guide for Public, 2023)

***Metadata***

On the webpage containing the links to the PDF and xlsx documents, there is some metadata including “Data Type: Nutrition Assistance Programs Keydata Release”, the size of each document, (525.38 KB for xlsx, 194.25 KB for pdf), and the last time the page was updated (January 12, 2024). In the documents themselves, there is a first title page/sheet, which includes the “origin” (Generated from National Data Bank Version 8.2 PUBLIC on 01/12/2024), a list of affiliated offices (Program Data Branch, Budget Division, Financial Management, Food and Nutrition Service, USDA) and a note that states: “This report is based in part on preliminary data submitted by various reporting agencies. Users should anticipate changes in future reports as reporting agencies finalize data.” There is also a contact for questions referring to the “data administrator, Budget Division 305-2189”. The second sheet or page includes a table of contents that lists acronyms for all the programs and associated tables (ex: CCCDCH-M1 is Child and Adult Care Food Program -- Child Care Type of Meal Served: Homes & Centers) and the letter definitions (ex: $ = Costs, P = Participation, etc). The footnotes on the sheets also include various details such as where the data came from, if data is combined, exceptions for certain program data that aren’t comparable, if data is incomplete, etc.

The metadata is fairly comprehensive with the table of contents and title page. However, the original source of the data (National Data Bank Version 8.1) is not cited with a link or described. This may be confidential data, but further elaboration on this key point of the origin of the data would be prudent. More metadata on the webpage would be useful for those previewing or browsing data so that users don’t have to open the document to see the basic summary. A summary or description of the data would be helpful.

As far as I could tell with little familiarity with metadata standards, there was not a metadata standard used here. It didn’t seem to match the DDI, FHIR, or ODM models listed under “Health Policies” in the metadata standards list. [(*Health Policy – Metadata Standards Catalog*, n.d.)](https://www.zotero.org/google-docs/?XyeiKW)

***Potential Improvements***

The main area this data is lacking is discoverability. I realized I wanted to do food program data and couldn’t find this data set through Figshare or Data.gov. I ended up finding it through Google and the USDA website, so it is currently not easily discoverable in the repositories. Notably, the Summary of Annual Data and Summary of Major Programs for Latest Available Month both have broken links, so even if someone can discover or find that data they can’t access it. I originally wanted to use those but couldn’t, so I deferred to the monthly Keydata that wasn’t broken. More descriptive titles, and tagging the relevant programs that are included in the report would also help with discoverability.

For users unfamiliar with this data, they do a good job giving notes on acronyms and definitions in the table of contents. The performance report may be more useful for users who are not familiar or need broader summaries. It would be helpful to include links to original sources of data in the footnotes or as a bibliography with links.

***Publications***

This is the newest monthly dataset available and was last updated on January 12, so there has not been time for publications about this particular month. However, the wider annual data is used by researchers, for projects like the “Trends in Supplemental Nutrition Assistance Program Participation Rates: Fiscal Year 2016 to Fiscal Year 2020”. [(Vigil, n.d.)](https://www.zotero.org/google-docs/?Jrbcca) One great thing about the USDA Data and Research pages where the datasets are housed is that there is a “Research and Analysis” page that shows research published or funded by the USDA utilizing data like this. Navigation is easy on this page, and a user can also narrow down the research by topic, like “Participation Rates” or “Food Security”

The name of this dataset makes it hard to search, but I tried “Nutrition Assistance Programs Keydata Release October 2023” which yielded no citations on Google Scholar or UW library’s search. This is unsurprising since it was published on Jan 10.

**Repository Profile**

Repository: Global Health Data Exchange: <https://ghdx.healthdata.org/>

I chose my repository, the Global Health Data Exchange (GHDx), as the most appropriate place for my dataset on USDA Nutrition Assistance Programs. [(*Global Health Data Exchange | GHDx*, n.d.)](https://www.zotero.org/google-docs/?thneJY) I found this repository by browsing re3data for the “Public Health, Health Services Research, Social Medicine” category. [(Re3data.Org, 2013)](https://www.zotero.org/google-docs/?jsiQ8b) Upon discovering this repository, I noted that it encompassed subcategories comparable to my dataset, such as food expenditures, financial assistance for health, and public social assistance. [(*Keywords | GHDx*, n.d.)](https://www.zotero.org/google-docs/?LnnCHS)

Under their “About Data Types” page, they list the data types that they categorize by. My dataset best falls under “Administrative data”, as defined by them as “data from the records maintained by agencies, institutions, commercial entities, and governments, where the records are used for administrative purposes or for providing services.” This dataset may also fall under the data type of “Financial record”, which is “data from the formal records of the financial activities of an entity (or person)”. [(*About Data Types | GHDx*, n.d.)](https://www.zotero.org/google-docs/?Rchyxr) I also confirmed that they did not already have my dataset, the October 2023 Keydata Report for Nutrition Assistance Programs. It was a bonus that the UW Institute for Health Metrics and Evaluation maintains this repository.

According to the re3data page for the GHDx, data submissions are limited to “other” and “institutional membership”. In other words, the data submission is not open to anybody. There are no parts of the website that I could find that show how to submit data. Rather, they have a section called “About our Information Sources”, where they credit data providers such as those listed on their “Data Sites We Love” page, and other organizations that they often get data from including IPUMS, ICPSR, and the World Bank (who they particularly credit for sharing extensive documentation). [(*About Our Information Sources | GHDx*, n.d.)](https://www.zotero.org/google-docs/?rzW4Cw) From the sites listed on these two pages, we can deduce that GHDx have ongoing relationships with large government organizations or non-profits who contribute their data. They do not currently have the USDA listed as a contributing entity, but it is in line with other government health organizations they do currently accept data from like the CDC and the U.S. Department of Human Services. The collection scope includes a range of global health metrics, including census data, disease registries, environmental monitoring, geospatial data, legislation, and survey data, to name a few of their Data Types. [(*About Data Types | GHDx*, n.d.)](https://www.zotero.org/google-docs/?XzvKVn) Thus, their scope is wide but unified by the common theme of global health information.

The GHDx has different kinds of data available with differing access levels, which they detail on their “About Data Availability” page. [(*About Data Availability | GHDx*, n.d.)](https://www.zotero.org/google-docs/?lESiOu) Depending on the type of data (detailed on each page with a red button link and label), different people can access it. “Microdata access: Download” are “available for download either directly or with a simple registration process that can be completed in one day or less”, the “Microdata access: Register” are “available for download with a registration process that takes longer than a day”, “Microdata access: Request” are “available by special request to the providing organization” but there may be a fee, and the “Tabulations only” datasets are “available either in a standalone file or in a report”. Users can filter their results according to the availability category.

The registration processes noted above are not for GHDx itself but rather for the institutions providing the resources. For example, ICSPR was listed under a dataset that was “Microdata: Register”, and that link took me to ICSPR where I would need to make an account or sign in through institutional access. Likewise for a dataset categorized as “Microdata access: Request”, one would have to request through the organization.

There are different access mechanisms according to the above access permissions. However, in terms of access mechanisms within these categories, it is dependent on the linked entity’s options and formatting. The only differentiation they show on GHDx is for tabulations which are only available as standalone files or in a report.

GHDx provides information on metadata standards on their “About Dataset Records” page. [(*About Dataset Records | GHDx*, n.d.)](https://www.zotero.org/google-docs/?eOSnq7) They do not cite a specific metadata standard but give information on giving unique citations and differentiation between datasets. For example, their title standardization is Geography - Name - Years.

I am drawing my DIP information from the dataset record page examples, as there are no pages describing this. Because I cannot see what is required in a submission package by the submitting party, this is mostly a deduction. The titles are created by the GHDx in the above format. Other metadata differs depending on the dataset, indicating that some submitters do not include all components and thus not all elements are required. Most include the title, provider, geography, coverage type, time period covered, series or system, data type, summary, and keywords. One of the less frequent components is a summary, likely because it would be voluntarily submitted. Each record has a red square button that directs the user to the dataset and is labeled according to access type. The records also include a Citation tab, which includes contributors and suggested citation formatting.

**Recommended Data Citation**

Program Data Branch, U.S. Department of Agriculture. (2024). October 2023 Keydata Report. U.S. Department of Agriculture Food and Nutrition Service. *doi*

**Consideration for Long-Term Preservation**

These file formats (Excel and PDF) are both well-maintained and popular file types. Excel has been maintained since 1985 and PDF has been maintained since 1993, so they are likely to remain accessible in the long term. The PDF version is especially useful in case future versions of Excel are not compatible with this file.

**Statement on Copyright License**

\*this seems to be in Module 9 for next week so will wait to do that module before entering info here

**Statement on Human Subject Considerations**

The data does not have personally identifiable data about people, and thus there was no reason to adjust the data for privacy or ethical reasons.

[*About Data Availability | GHDx*. (n.d.). Retrieved February 9, 2024, from](https://www.zotero.org/google-docs/?Sxi9xB) <https://ghdx.healthdata.org/about-ghdx/about-data-availability>

[*About Data Types | GHDx*. (n.d.). Retrieved February 9, 2024, from](https://www.zotero.org/google-docs/?Sxi9xB) <https://ghdx.healthdata.org/about-ghdx/data-type-definitions>

[*About Dataset Records | GHDx*. (n.d.). Retrieved February 9, 2024, from](https://www.zotero.org/google-docs/?Sxi9xB) <https://ghdx.healthdata.org/about-ghdx/ghdx-records-explained>

[*About our Information Sources | GHDx*. (n.d.). Retrieved February 9, 2024, from](https://www.zotero.org/google-docs/?Sxi9xB) <https://ghdx.healthdata.org/about-ghdx/our-information-sources>

[*Calculating the Supplmental Nutrition Assistance Program (SNAP) Program Access Index: A Step-by-Step Guide for 2013*. (2015). USDA Food and Nutrition Service.](https://www.zotero.org/google-docs/?Sxi9xB) <https://www.cwda.org/sites/main/files/file-attachments/pai2013.pdf?1455134856>

[*Data & Research | Food and Nutrition Service*. (n.d.). Retrieved January 26, 2024, from](https://www.zotero.org/google-docs/?Sxi9xB) <https://fns-prod.azureedge.us/data-research?f%5B0%5D=research_type%3A199>

[*Global Health Data Exchange | GHDx*. (n.d.). Retrieved February 9, 2024, from](https://www.zotero.org/google-docs/?Sxi9xB) <https://ghdx.healthdata.org/>

[*Grants | Food and Nutrition Service*. (n.d.). Retrieved January 26, 2024, from](https://www.zotero.org/google-docs/?Sxi9xB) <https://fns-prod.azureedge.us/fm/grant-opportunities>

[*Health policy – Metadata Standards Catalog*. (n.d.). Retrieved January 26, 2024, from](https://www.zotero.org/google-docs/?Sxi9xB) <https://rdamsc.bath.ac.uk/subject/Health%20policy>

[*Keywords | GHDx*. (n.d.). Retrieved February 9, 2024, from](https://www.zotero.org/google-docs/?Sxi9xB) <https://ghdx.healthdata.org/keywords>

[*October 2023 Keydata Report | Food and Nutrition Service*. (2024, January 12).](https://www.zotero.org/google-docs/?Sxi9xB) <https://fns-prod.azureedge.us/data/October-2023-keydata-report>

[Re3data.Org. (2013). *Global Health Data Exchange*. https://doi.org/10.17616/R3QK52](https://www.zotero.org/google-docs/?Sxi9xB)

[*Research & Analysis | Food and Nutrition Service*. (n.d.). Retrieved January 26, 2024, from](https://www.zotero.org/google-docs/?Sxi9xB) <https://fns-prod.azureedge.us/research-analysis>

[Vigil, A. (n.d.). *Trends in Supplemental Nutrition Assistance Program Participation Rates: Fiscal Year 2016 to Fiscal Year 2020*.](https://www.zotero.org/google-docs/?Sxi9xB)