

Data-usage descriptors as search metadata: the case of food security data and the National Data Platform (2015-2025)

Lauren Chenarides

Colorado State University

Rafael Ladislau

RL Desenvolvimento de Sistemas LTDA

Manish Parashar

University of Utah

Simon Porter

Digital Science (United Kingdom)

Julia Lane

`julia.lane@nyu.edu`

New York University

Article

Keywords:

Posted Date: March 4th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-8569040/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Data-usage descriptors as search metadata: the case of food security data and the National Data Platform (2015-2025)

Lauren Chenarides¹, Rafael Ladislau², Manish Parashar³, Simon Porter⁴, Julia Lane⁵

Article Type: **Article**

Abstract

Scientific data is a critical input into scientific research. Yet the research data landscape is constantly changing as new datasets emerge, others are retired, or some disappear altogether. Data-usage descriptors can substantially advance research productivity by reducing the time that researchers spend finding new and relevant datasets in their research field. This paper describes how to generate data usage descriptors by finding how datasets are used in publications and then linking the dataset information to the publication metadata. It also shows how usage descriptors can be used to find other related datasets and their usage. It concludes by arguing that the approach represents a critical piece of foundational infrastructure that could be deployed in repositories as part of a referenceable, navigable, and contextual data framework. This article contains a reproducible workflow for constructing data-usage descriptors, based on analyzing the full text of publications in the Dimensions database. The illustrative use case is research on food security. The illustrative repository is the National Data Platform.

1 – Assistant Professor, Department of Agricultural and Resource Economics, Colorado State University, Lauren.Chenarides@colostate.edu

2 – CEO, RL Desenvolvimento de Sistemas LTDA, rafa.ladis@gmail.com

3 – Presidential Professor and Chief AI Officer, Kahlert School of Computing, University of Utah, manish.parashar@utah.edu

4 – VP of Research Futures, Digital Science, s.porter@digital-science.com

5 – Professor Emerita, Robert F. Wagner Graduate School of Public Service, New York University, jil4@nyu.edu

Introduction

Scientific data is a critical input into scientific research. The research data landscape is constantly changing as new datasets emerge, others are retired, or some disappear altogether. New data can arise for many reasons, including deliberate investments in specific fields such as the Protein Data Bank (Heikkilä, October 15, 2024), ImageNet (Li, 2023), or the Sloan Digital Sky Survey (Szalay et al., 1999), the emergence of superior data collection technologies (Meyer et al., 2015), or the emergency of new scientific challenges, such as the novel coronavirus (Shuja et al., 2021). Datasets can also disappear as a result of deliberate disinvestment decisions, such as the 2023 cuts to the US National Ecological Observatory Network (Mervis, 2023), or lack of use (Gibney & Van Noorden, 2013), while others are inconsistently collected due to funding and implementation challenges as in the case of the Survey of Inmates in Local Jails (Statistics, 2024). In recent months, US federal datasets have been quite rapidly retired (Skelley, 2025), affecting both the productivity of both scientists and policy researchers in a number of fields.

Regardless of the cause, the constant churn in scientific datasets creates both a fundamental challenge and opportunity for empirical science. Researchers in fields where datasets are disappearing do not have a systematic way of finding other relevant data sources, or signals of how they are used. Similarly, researchers in fields where datasets are rapidly appearing do not have a systematic way of finding what they are used for and who are the experts. The mechanisms that have evolved to find the written scientific output captured in journal publications – such as the authors, their institutional affiliations, the journals, and the associated citations – do not exist for data in a consistent way.

This paper outlines a mechanism to develop usage descriptors for data that parallels and is derived from publications. Data-usage descriptors – or structured metadata that link datasets to publications, citations, authors, institutions, and journals – could be intentionally developed as a form of data about the research system that enhances the study of science and innovation. If fully developed within data repositories or other data infrastructures, they could complement familiar publication indicators such as citation counts or grant data by enabling repositories to characterize how underlying datasets are used as inputs into scientific work (Big Data Interagency Working Group, 2022; Parashar, 2024). They could provide evidence to the funders of data about which parts of the data infrastructure are heavily relied upon, which communities depend on them, and where the system may be vulnerable when datasets are introduced, change, or are discontinued. They could also help identify dataset experts and reward them for their contribution to dataset development in a manner similar to Project CRediT (Brand et al., 2015) and create a more productive system.

This paper demonstrates the development and potential value of data-usage descriptors for an illustrative use case: food security research. The use case is an important one: empirical research on food security is used to inform decisions that affect almost 50 million Americans, including 14 million children, particularly those who rely on public benefit programs (Rabbitt et al., 2025). Much of that research relies on a major US Department of Agriculture dataset – the Household Food Security Survey Module (HFSSM) – which has served as the official measure of food insecurity for three decades and as the basis for thousands of studies on household well-being and nutrition policy (Carlson et al., 1999; Rabbitt et al., 2023) – and which has been targeted for elimination (US Department of Agriculture, 2025).

It also describes how data-usage descriptors have been deployed for the case of food security as part of a new data repository¹ funded by the National AI Research Resource – the National Data Platform (Parashar & Altintas, 2023). That platform is designed to support data user needs for data-driven scientific workflows, such as ease of access and use, time to science, and energy/environmental impact. (Parashar, 2025; Parashar & Altintas, 2023)

This paper proposes a method for developing data-usage descriptors for a defined list of datasets in the case of food security. It also outlines how the methods can be expanded to both find alternative data sources and generate data-usage descriptors for those sources. It provides a code and workflow that can be replicated for any field with an identifiable core set of datasets, such as climate science or labor economics. That same workflow could be used to trace how those datasets appear in the literature, identify complementary or substitute data sources, and map research communities that form around them. It concludes with a description of how the approach could be institutionalized as part of critical data infrastructures which are designed to host new types of data like the National Data Platform.

Methods

Context

This article builds on a successful effort to developing data usage statistics based on finding references to datasets in the full text of scientific publications in what was initially called the Rich Context project (Lane et al., 2022; Lane et al., 2020), and subsequently the Democratizing Data project (Meng, 2024). A prototype system to collect usage measures for eleven USDA datasets,² including the HFSSM, was funded as part of a response to the 2018 Foundations of Evidence-based Policymaking Act (115th Congress Public Law 115–435, 2018), which required agencies to publish on their websites, on a regular basis (but not less than annually), information on the usage of public data assets by non-government users.

The potential value for using the system and its usage statistics to document the effect of discontinuing the HFSSM was highlighted in a LinkedIn post by a high profile food security researcher in October 2025, which reported not only the number of journal articles, but the citations, the authors, and the number of institutions – information which would not have been able to be found using manual approaches. (Masters, 2025).

¹ <https://food.di.ndp.utah.edu/>

² The eleven datasets included: Census of Agriculture, Current Population Survey Food Security Supplement (CPS FSS), Household Food Security Survey Module (HFSSM), Food Acquisition Purchase Survey (FoodAPS), Agricultural Resource Management Survey (ARMS), Farm to School Census, Food Access Research Atlas (FARA), Circana (formerly IRI) Infoscan data, Local Food Marketing Practices Survey, Quarterly Food-at-Home Price Database, Tenure, Ownership, and Transition of Agricultural Land (TOTAL) Survey, and are cited in Appendix A.2. These core USDA data are derived from survey, administrative, and/or proprietary sources. They are widely used in research on food systems, most often in studies of food security, but also in research on production, consumption, access, and market behavior. They were used as a “seed set” for demonstrating how a usage-descriptor workflow can be constructed and scaled. The early project used Elsevier’s Scopus dataset, and the approach featured different coverage decisions, used different methodology, and applied different selection criteria. The results from that project consequently differ from those reported in this paper.

Data

The project experimented with a variety of different scientific databases (Hausen & Azarboyad, 2024; Lane et al., 2020; Potok, 2022). The Dimensions database (Hook et al., 2018) was chosen to develop data usage descriptors, not only because of its breadth of coverage and the quality of its curation, but also the openness to scientific collaboration, and the shared scientific vision of the Digital Science team. The Dimensions database contains structured metadata linking more than 160 million publications to datasets, funders, institutions, and research topics. It provides full-text search capabilities through an application programming interface (API).³ The access was also researcher friendly – the team was provided with direct access to the Dimensions corpus via Google BigQuery, which allowed for automated and reproducible data retrieval. Where data analysis required access to full text beyond the scope of distribution licenses, this work was handled by Digital Science team members.

Methods

The workflow consists of two stages. The first stage developed data usage descriptors for the USDA datasets relevant to food security; the second stage identified other, non-USDA datasets used in food security research and developing data usage descriptors for those data.

Stage One includes three primary steps. The first two steps rely on a string-search approach to search the Dimensions corpus to identify publications that reference USDA datasets: the “seed set.” The third step determines whether the dataset mention in each identified publication reflects an actual use of the data, using a more computationally intensive process based on full text and a multi-agent LLM engine.

The result of these three steps is a set of data usage statistics for each seed set dataset that reports the number of publications that made use of the data, the authors and institutions of those publications, the number of citations, and the journals in which each publication appeared.

Stage Two makes use of the first stage corpus to identify related datasets outside of the seed set and develops the same set of data usage statistics for them.

Each step for stages one and two is described in detail below, and illustrated conceptually in Figure 1.

³ The Dimensions-derived dataset was validated against two additional bibliographic sources, Scopus and OpenAlex, using six datasets: ARMS, Census of Agriculture, Food Access Research Atlas, FoodAPS, HFSSM, and RUCC. This approach and the results are presented in Appendix A.1.

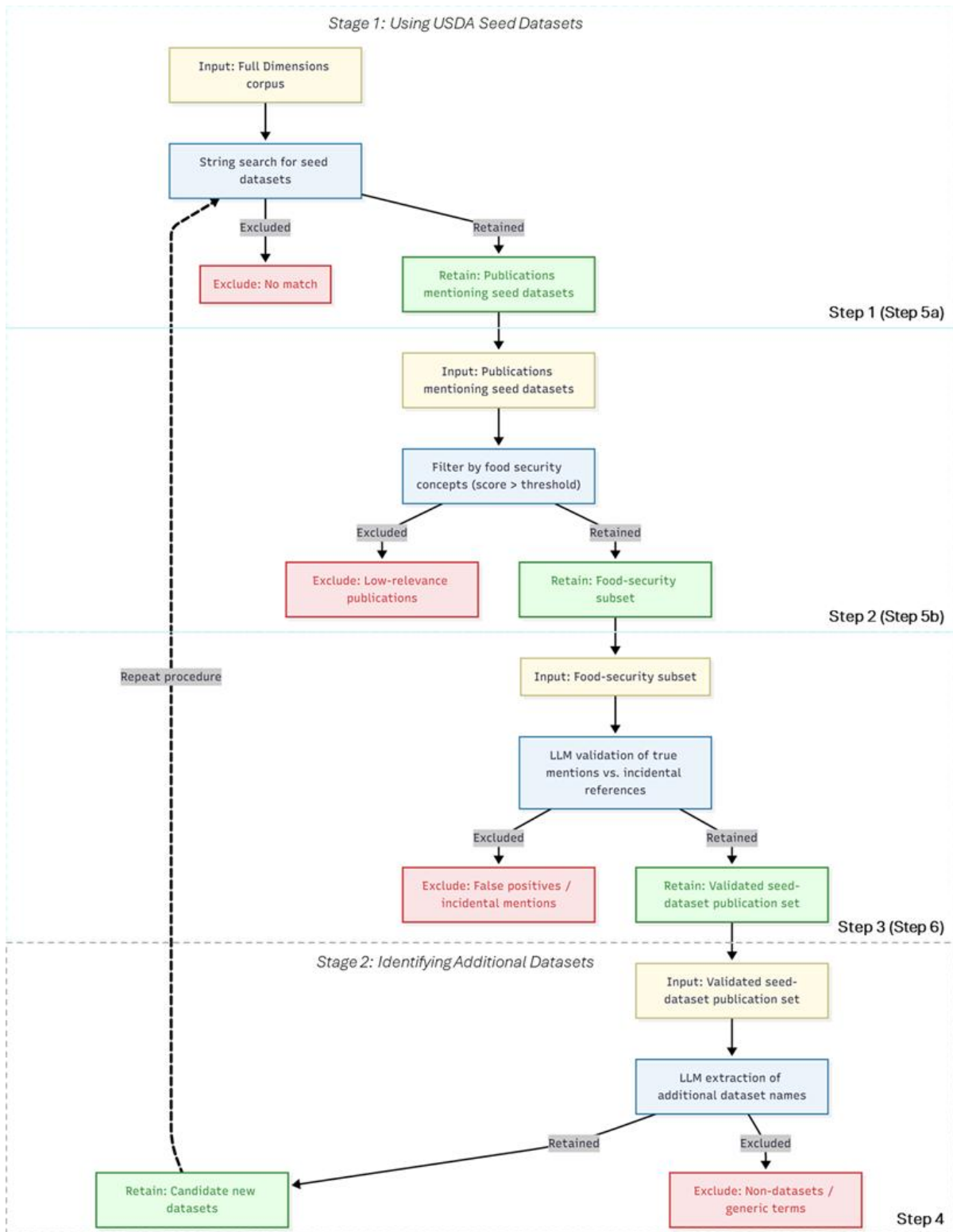


Figure 1: Mermaid diagram depicting the workflow. Designed using Mermaidchart.com. Stage One (steps 1-3) develops data-usage descriptors for the USDA seed datasets. Stage Two (steps 4-6) develops data-usage descriptors for newly identified datasets. Steps 5a, 5b, and 6 mimic steps 1, 2, and 3, in stage one.

Stage 1: Developing data-usage descriptors for USDA datasets

Step 1: The first step was to identify publications that reference the initial set of eleven datasets⁴ which served as the seed group for this workflow.

The entire Dimensions corpus consists of **159 million** publications. This can be queried using Boolean operators using Dimensions’ Domain Specific Language (DSL). Searches were restricted to publications published between 2015 and 2025 and authored by at least one researcher affiliated with a US institution based on the Dimensions address mapping approach (Guerrero-Bote et al., 2021), yielding a search corpus of **7,473,055** publications represented by distinct DOIs.⁵ The query matched the official dataset names and their commonly used aliases in the title, abstract, or full text. They were executed in Python using the dimcli client, and the results were returned as structured JSON objects.⁶

This initial step resulted in **11,915** publications that mentioned at least one of the seed USDA datasets, noting that these mentions could include both substantive uses and incidental references. The API was used to pull key fields: the publication title, DOI, journal name, abstract, publication year, citation counts, subject classifications, authors, and institutional affiliations. This information forms the foundation of the data-usage descriptors, which indicate who uses a given dataset, in which research areas, and in what publication outlets.

Step 2: The next step was to narrow down the publications that mentioned the USDA seed datasets to only those relevant to food security. In Dimensions, “concepts” describe the main topics of a publication and are derived from the full text using machine learning.⁷ Each publication can be linked to multiple concepts, each of which is assigned a relevance score between 0 and 1 indicating how closely the concept relates to the publication’s content. The search used two concepts—“food security” and “food insecurity”—to identify relevant publications. Of the **11,915** publications derived from step one, **1,596** publications are associated with “food security” or “food insecurity” concepts, but many are only loosely linked, and these would have a low relevance score (Table 1).⁸

⁴ Researchers often reference datasets inconsistently, using acronyms, abbreviations, alternate spellings, or related URLs. We developed a structured list of dataset-alias pairs (“dyads”) to map each dataset’s official title to common variants. For example, the “Food Acquisition and Purchase Survey” was linked with the alias “FoodAPS.”

⁵ The US restriction was because of the US focus; additional restrictions included institutional identifiers such as “USDA,” “NASS,” or “U.S. Department of Agriculture.” These terms were grouped with OR within each category and joined with AND across categories. The approach could of course be global, since Dimensions covers publications from a wide variety of international sources.

⁶ Each full-text search took about five minutes. Processing a filtered publication corpus of n=2,321 on a Mac Studio (M4 Max, 128 GB RAM) required roughly eight continuous days of runtime, which implies that applying the same deep analysis to the full Dimensions corpus of 7.4 million publications without prior filtering would be computationally infeasible.

⁷ Refer to <https://api-lab.dimensions.ai/cookbooks/1-getting-started/7-Working-with-concepts.html>; Future work could include using LLMs to identify concepts (see <https://www.dimensions.ai/news/introducing-dimensions-research-gpt-and-dimensions-research-gpt-enterprise/>).

⁸ From the 7.4 million publications in our starting corpus (English-language, peer-reviewed articles published between 2015 and 2025 with at least one U.S.-affiliated author), 15,531 publications included “food security” or

For the purpose of this illustrative use case, the publications with a food security relevance score greater than 0.6 were selected, which had the additional advantage of ensuring that the sample size was computationally manageable for the LLM engine. Under this threshold, the filtered dataset contained approximately **1,014** publications.

Table 1. Frequency table of number of publications by food security relevance score

Score Threshold	Count	Percentage
>= 0.1	1,596	100.00%
>= 0.2	1,595	99.94%
>= 0.3	1,592	99.75%
>= 0.4	1,492	93.48%
>= 0.5	1,350	84.59%
>= 0.6	1,014	63.53%
>= 0.7	487	30.51%
>= 0.8	13	0.81%
>= 0.9	0	0.00%
>= 1.0	0	0.00%

Step 3: String search alone cannot distinguish substantive use from incidental mentions, a limitation raised during workshops with food security researchers who noted that such distinctions are important for empirical work and for identifying data user communities (Chenarides, 2025).

The next step was therefore to review each of the 1,104 potential matches to determine whether a dataset was actually used or simply referenced. A dataset review agent was developed within a multi-agent LLM engine that supplanted the manual validation approach used in early stages of the project (Emecz et al., 2024). The LLM had a number of advantages over manual approaches, most obviously that it was lower cost, more timely, more scalable and provided rich contextual information.

For each potential match, the engine received a prompt indicating that a publication mentioned a specific dataset identified through the string-search step, together with the full text of the publication provided by Dimensions. Because the string search does not incorporate context, the review engine assessed both whether the correct dataset was mentioned in the text, and whether the dataset was substantively mentioned. Two scores were produced. The first was a *dataset-mention* confidence score with a scale ranging from 0 to 10. Higher values indicated greater confidence that the reference pertained to the relevant dataset (for example, confirming that a mention of the “Census of Agriculture” referred to the US Census of Agriculture). The second was a *dataset-use* confidence score also scaled from 0 to 10, with higher values indicating greater

“food insecurity” as a Dimensions concept (at any relevance score), before applying the additional restriction that they also mention at least one USDA dataset.

confidence that the referenced dataset was used in the analysis of the publication rather than simply described. The engine based the dataset use score on six criteria.⁹

For the purposes of this illustrative use case, the corpus was subset to include those publications with both a confidence score exceeding 6 that the dataset mention referred to the relevant USDA dataset, and with a confidence score exceeding 6 that the dataset mention reflected actual use. The result for the number of publications that reported a dataset-mention confidence score over 6 was **859** publications. The result for the number of publications that reported a dataset-use confidence score over 6 was **602** publications. The identification of publications from the Dimensions database reflecting steps one through three (stage one) is depicted in Figure 2.

⁹ The six criteria include: 1. The paper performs analysis, modeling, or experiments on the dataset; 2. Results, findings, or insights are derived from the dataset; 3. The dataset is central to the research methodology; 4. Specific characteristics, variables, or samples from the dataset are discussed; 5. The dataset is used for data processing, statistical analysis, or machine learning; 6. The research methodology explicitly describes using this dataset.

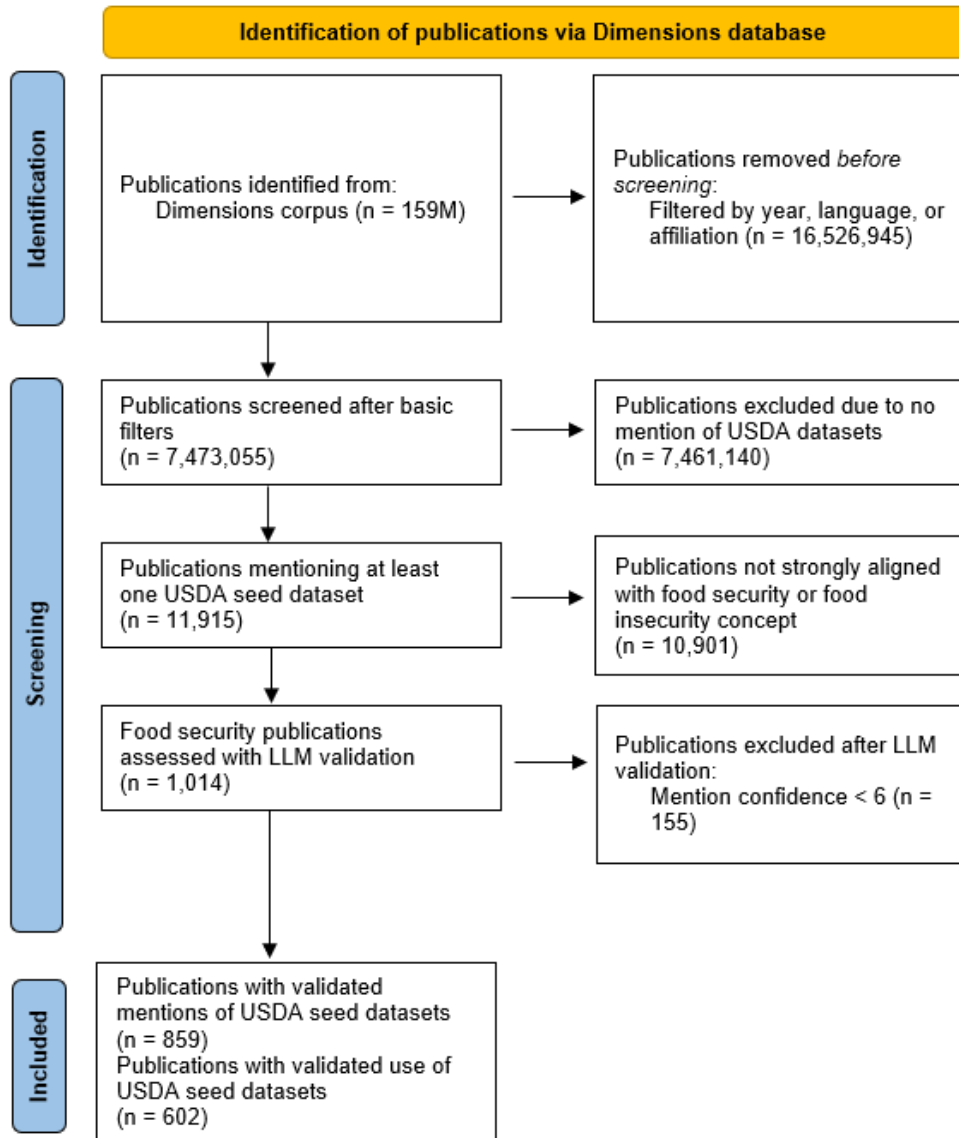


Figure 2: PRISMA diagram depicting identification of validated publications mentioning USDA seed datasets from Stage One

Stage 2: Identifying additional datasets used in food security research

Step 4: The next step was to identify a broader set of other potential data sources. This required a LLM-assisted extraction workflow designed to process full text and generate structured metadata describing mentions of other datasets. It is computationally infeasible to apply the LLM model to the entire food security dataset, so for expositional purpose, the model was applied to the corpus of **859** food-security publications with validated mentions of any of the eleven seed datasets from step 3.

The model implemented in this step identified 774 additional dataset names from each publication’s full text that were not part of the initial eleven USDA datasets.¹⁰ These extracted names ranged from broad references (e.g., “US Census data”) to specific datasets (e.g., “Feeding America Map the Meal Gap”). Each extracted name was paired with an evidence snippet indicating how and where it appeared in the publication. Because the raw output contained spelling variations and semantically similar dataset names, a string similarity grouping procedure was used to cluster together likely comparable dataset names, resulting in a consolidated list of approximately 77 candidate datasets.¹¹

These 77 candidate datasets were then ranked based on three criteria: (1) an agent-generated dataset-mention confidence score above 8 (on a 0–10 scale), (2) frequent appearance across the validated food security corpus, and (3) manual confirmation that the dataset is an actual dataset and not a questionnaire or non-dataset artifact. This resulted in 18 additional datasets from the validated food security publication corpus that are considered a new set of seed datasets.¹² The 18 additional datasets are listed in Appendix A.2, along with the agency or organization responsible for their collection. They include major data products from the US Census Bureau, the Centers for Disease Control and Prevention, the US Bureau of Labor Statistics, universities, as well as additional USDA datasets.

Step 5: The next step was to apply the same workflow used for the eleven USDA seed datasets to the 18 additional datasets and their aliases,¹³ namely (1) a string-search stage to identify candidate publications (steps 5a and 5b).

Step 5a: The string search described in step 1 was deployed on the full Dimensions publications corpus for the 18 newly discovered datasets. As with the USDA queries, these searches used the Dimensions DSL and were restricted to English-language, peer-reviewed publications from 2015–2025 with at least one U.S.-affiliated author. This step produced a combined set of 233,446 publications that mentioned at least one of the 18 datasets. Because this stage does not evaluate substantive use, the resulting counts include incidental mentions and references appearing in the text.

Step 5b: To maintain comparability with the earlier stages, the publication set was filtered to include only those for which the Dimensions concepts “food security” or “food insecurity” had a

¹⁰ The specific approach was to apply a string similarity grouping procedure using the RapidFuzz library's Levenshtein ratio to cluster likely comparable dataset names. First, the extracted names matching the initial 11 USDA datasets were removed. Then, pairwise similarity was computed for all remaining names, and a greedy clique-based algorithm grouped datasets where all members exceeded a 70% similarity threshold. Canonical names were selected based on highest average confidence scores. This reduced the initial set of approximately 2,600 extracted dataset mentions (or 1,809 after pre-filtering) to a consolidated list of 774 candidate datasets. The top 80 candidate clusters (ranked by member count) were then manually reviewed, from which 19 discovered datasets were selected for validation. These datasets and their aliases were subsequently used to query the Dimensions database for additional publications mentioning these newly identified datasets.

¹¹ The LLM also generated a report describing how datasets were combined within individual studies. These results are presented in Appendix A.3.

¹² Further details about the manual validation employed are described in Appendix A.4.

¹³ The same approach was used to identify aliases as for the USDA seed datasets: drawing on official dataset titles, common abbreviations, and naming variants appearing in the text. The additional datasets identified in this step are cited in Appendix A.2.

relevance score greater than 0.6. Applying this threshold reduced the candidate set to 2,237 publications, a number that remained feasible for the full-text review. This relevance-based filtering follows the same logic as in step 2. The number of validated publications mentioning additional datasets from stage two is presented in Table 2.

Step 6: The reviewed publication corpus introduced publications, authors, institutions, citations, and journals outlets that did not appear in the USDA-focused corpus. As before, all were deduplicated using Dimensions identifiers, such as publication ID for publications, ORCIDs for authors, RORs for institutions, and ISSNs for journals.

Results

Stage 1 Results: Data-usage descriptors for USDA food security datasets

The string search for the eleven USDA seed datasets identified 11,915 publications that mentioned at least one of these datasets across 2,288 journals, 32,417 authors, and 3,166 institutions. Applying the food security and food insecurity concepts with a relevance threshold greater than 0.6 reduced this set to 1,014 publications. These publications span 518 journals, involve 6,223 authors affiliated with 1,011 institutions, and account for 32,440 citations, and each publication mentioned at least one of the USDA seed datasets.

These steps produce a publication-level corpus of food-security research that directly links 2,480 authors affiliated with 561 institutions, and account for 13,006 citations, and each publication used at least one of the USDA seed datasets.

Stage 2 Results: Data-usage descriptors for additional food security datasets

The string search for the 18 additional datasets identified 233,446 publications that mentioned at least one of these datasets across 14,180 journals, 796,529 authors, and 19,871 institutions. Applying the food security and food insecurity concepts with a relevance threshold greater than 0.6 reduced this set to 2,237 publications. These publications appear across 657 journals, authored by 7,256 individuals affiliated with 1,322 institutions, generating 48,635 citations, and each publication mentioned at least one of the additional datasets.

Beyond the sheer counts of publications, when the corpus is expanded to reflect the validated food-security publications mentioning the 18 additional datasets, 4,017 additional authors affiliated with 571 institutions are identified, and, together, these publications have been cited 27,726 times.

Table 2. Process of identifying validated publications mentioning additional datasets from Stage Two

STEP	CORPUS CLASSIFICATION	COUNT	DESCRIPTION
IDENTIFICATION	Entire Dimensions Corpus (without any filters)	159,000,000	Number of publications in the entire Dimensions corpus, before applying the filters below, per dimensions.ai
SCREENING	Dimensions Corpus Articles published between 2015 and 2025 with at least one author affiliated with a US institution	7,473,055	Number of publications after applying the first set of filters to the screening process
SCREENING	Publication count mentioning at least one of the 18 “newly discovered” datasets	233,446	Number of publications mentioned at least one of the additional datasets, before applying the “food security” or “food insecurity” concept (with any score)
SCREENING	Publication count food security or food insecurity concept	3,207	Number of publications after applying a minimum relevance score of 0.1
SCREENING	Publication count food security or food insecurity concept ≥ 0.6	2,237	Number of publications after applying a minimum relevance score of 0.6

Data-Usage Descriptors and Structured Data Discovery

The food security data-usage descriptors, as well as the reproducible workflow for constructing data-usage descriptors presented in this paper, are being used to support a usage-based contextual data discovery service within the National Data Platform (NDP).¹⁴ The NDP aims to transform data accessibility and usability for researchers across diverse scientific domains by providing a

¹⁴ <https://nationaldatapatform.org/>

federated, extensible data ecosystem of data resources and services. It provides a cohesive platform that supports comprehensive data services for ingestion, indexing, curation, and analysis. Figure 3 shows how the usage statistics can be portrayed (for a smaller set of years) on the NDP.

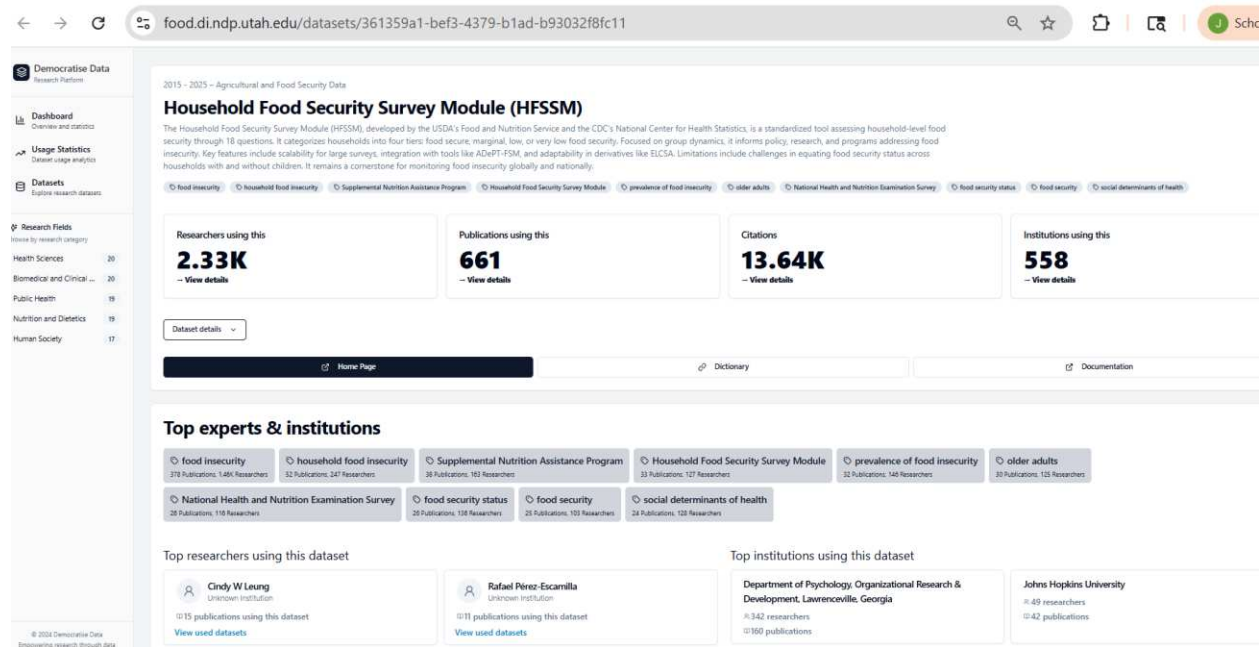


Figure 3: The National Data Platform and HFSSM usage statistics

The data-usage descriptors are integral to a multidimensional framework for data trust currently under development. Specifically, the framework combines data accessibility—its availability and the quality of its provenance metadata and governance processes—with context-specific trust indicators based on usage and end-user feedback verification. NDP hosts diverse data from a range of sources. These include domain data repositories (e.g., the NIH MIDRC Data Commons¹⁵), streaming data from instruments and observatories (e.g., data from the NSF EarthScope Consortium¹⁶ and the NOAA and NASA GOES Satellite Network¹⁷). NDP also integrates end-to-end cyberinfrastructure for real-time, data-driven simulation, prediction, and visualization of wildfire behavior (WIFIRE), which can be used to model and predict wildfire spread (Parashar & Altintas, 2023). Moving forward, NDP will continue to expand its data resource to include a range of emerging data sources such as streaming data from Internet of Things networks, smart cities, low-orbit satellites, and social media, and support workflows that combine these data streams with more traditional data sources to support data-driven understanding and decision making. A major advantage is that because the workflow has been rigorously developed to be reproducible, it is relatively low-cost to implement and scalable to multiple domains.

¹⁵ MIDRC Data Commons, <https://data.midrc.org/>.

¹⁶ The EarthScope Consortium, <https://www.earthscope.org/>.

¹⁷ GOES Satellite Network, <https://science.nasa.gov/mission/goes/>.

Discussion

The data curation workflow presented in this article shows both how data-usage descriptors can be developed and how the methodology can be used to learn about additional datasets and their usage, regardless of domain.¹⁸ The paper is illustrative in nature, and represents a conservative view of the use of datasets in food security research rather than a definitive count of all uses. Both the corpus and the methodology could be enhanced and expanded with the engagement of the science and innovation community. There are at least three ways in which that engagement could advance both science and innovation.

First, data-usage descriptors could be used to advance understanding of datasets that are used in scientific research areas. The descriptors can not only be used to provide new information about existing data usage, but the methodology can be deployed to identify additional datasets – and their usage in the same field. Data creators that wanted to get more credit for their data being used could make their data more findable. For fields affected by dataset retirements, this type of usage mapping points to potential substitute data sources and ways of identifying the communities that already work with them.

Second, the data-usage descriptors generated by this workflow could be used by science agencies, publishers, and research fields, to help reward and incentivize investments in data. In the food security case, data-usage descriptors support three related lines of analysis. They describe inputs to research by showing how often particular federally funded datasets appear in the literature, and which alternative datasets are drawn in from other domains, including labor statistics, public health data, or demographic data. They also characterize outputs and the evolution of activity over time, including trends in publication counts, citations, and journal outlets associated with specific datasets or groups of datasets. This information could help both inform funding decisions and reward the production and reuse of important datasets by researchers (Lane et al., 2024).

Finally, despite the importance of data for science, and in propelling scientific advances, data have not been treated as a “first class output.” This is, in part, because they do not have the institutional framework that has been developed around research papers. Data citations, metadata, classifications, persistence are all in the early stages of the formation of norms. Systems like the National Data Platform can be critical in providing norms that enable data to exist in a persistently referenceable, navigable and contextual framework (Parashar, 2025; Parashar & Altintas, 2023).

Simply put, data usage measures if broadly adopted by the scientific community and implemented by the data repositories that serve the community, could help inform the development of a system to navigate the ever-changing data ecosystem landscape.

¹⁸ The workflow, queries, and file structure can be reused to construct similar data-usage descriptors for other seed datasets or research domains. In areas where core datasets are at risk of retirement or where the data landscape is fragmented, this approach provides a practical way to document data use, identify gaps, and locate existing expertise for future work on science and innovation.

Data Availability

The data used in this research is available in a GitHub repository: <https://github.com/national-data-platform/food-security-usage-descriptors>. Accompanying data documentation is provided in the GitHub repository, as well as in Appendix A.5.

Code Availability

The code used in this research is also available in the GitHub repository: <https://github.com/national-data-platform/food-security-usage-descriptors>.

Acknowledgements

The authors acknowledge the substantive contributions of Daniel Hook, Nick Pallotta, Spiro Stefanou, Mark Denbaly, Nancy Potok, and May Aydin.

Author Contributions

L.C. contributed to the investigation, writing, reviewing and editing, visualization, validation, and project administration.

J.L. contributed to the conceptualization, methodology, writing of the original draft, reviewing and editing, project administration, supervision and funding acquisition.

R.L. contributed to the methodology, software, validation, formal analysis, investigation, data curation, writing – review and editing, and visualization.

M.P. contributed to the conceptualization, software, validation, writing and funding acquisition.

S.P. contributed to software, validation, resources, data curation, writing – reviewing and editing, project administration, supervision.

All authors reviewed the manuscript.

Competing Interests

None

Research Funding

This work was supported by gifts, grants and contracts from the US Department of Agriculture 58-5000-8-0053, the National Science Foundation contract 49100422C0028, grant 2333609, grant 2440195, Schmidt Sciences, Patrick J. McGovern Foundation, Overdeck Family Foundation, and the Alfred P. Sloan Foundation. Digital Science provided financial support to Rafael Ladislau.

References

- Foundations for Evidence-Based Policymaking Act of 2018, (2018).
- Big Data Interagency Working Group. (2022). *Big Data: Pioneering the Future of Federally Supported Data Repositories Workshop Report*. Retrieved from <https://www.nitrd.gov/pioneering-the-future-of-federally-supported-data-repositories/>
- Brand, A., Allen, L., Altman, M., Hlava, M., & Scott, J. (2015). Beyond authorship: Attribution, contribution, collaboration, and credit. *Learned Publishing*, 28(2).
- Carlson, S. J., Andrews, M. S., & Bickel, G. W. (1999). Measuring food insecurity and hunger in the United States: development of a national benchmark measure and prevalence estimates. *The Journal of nutrition*, 129(2), 510S-516S.
- Chenarides, L. (2025). Envisioning food and agricultural data for the future: Engaging communities to identify trusted data. In. <https://www.cfare.org/items/foodandagdata>: Council on Food, Agricultural & Resource Economics.
- Chenarides, L., Bryan, C., & Ladislau, R. (2025). *Methodology for comparing citation database coverage of dataset usage*. In https://laurenchenarides.github.io/data_usage_report/report.html
- Chenarides, L., Hanks, A. S., Berard, J., Carlson, A. C., Davis, G., & Finaret, A. B. (2025). A review of data linkages for policy-informing research in food and agricultural economics. *Food Policy*, 137, 102996.
- Emecz, A., Mitschang, A., Zdawczyk, C., Dahan, M., Baas, J., & Lemson, G. (2024). Turning visions into reality: Lessons learned from building a search and discovery platform. *Harvard Data Science Review*.
- Gibney, E., & Van Noorden, R. (2013). Scientists losing data at a rapid rate. *Nature*. <https://doi.org/10.1038/nature.2013.14416>
- Guerrero-Bote, V. P., Chinchilla-Rodríguez, Z., Mendoza, A., & de Moya-Anegón, F. (2021). Comparative analysis of the bibliographic data sources Dimensions and Scopus: An approach at the country and institutional levels. *Frontiers in Research Metrics and Analytics*, 5, 593494.
- Hausen, R., & Azaronyad, H. (2024). Finding the Data: An Ensemble Approach to Uncovering the Prevalence of Government-Funded Datasets.
- Heikkilä, M. (October 15, 2024). A data bottleneck is holding AI science back, says new Nobel winner. *MIT Technology Review*. <https://www.technologyreview.com/2024/10/15/1105533/a-data-bottleneck-is-holding-ai-science-back-says-new-nobel-winner/>

- Hook, D. W., Porter, S. J., & Herzog, C. (2018). Dimensions: building context for search and evaluation. *Frontiers in Research Metrics and Analytics*, 3, 23.
- Lane, J., Gimeno, E., Levitskaya, E., Zhang, Z., & Zigoni, A. (2022). Data inventories for the modern age? Using data science to open government data. *Harvard Data Science Review*, 4(2).
- Lane, J., Mulvany, I., & Nathan, P. (2020). Rich search and discovery for research datasets: Building the next generation of scholarly infrastructure. In: Sage London.
- Lane, J., Spector, A. Z., & Stebbins, M. (2024). An Invisible Hand for Creating Value from Data: The opportunities resulting from new legislation and technologies. *Harvard Data Science Review*.
- Li, F.-F. (2023). *The Worlds I See: Curiosity, exploration, and discovery at the dawn of AI*. Flatiron books: a moment of lift book.
- Masters, W. [<https://www.linkedin.com/in/wamasters/>]. (2025). <https://www.linkedin.com/feed/update/urn:li:ugcPost:7375859958065553408/>
- Meng, X.-L. (2024). Data Democratization: An Ecosystemic Contemplation and Coordination. *Harvard Data Science Review*.
- Mervis, J. (2023). Researchers protest end of NSF grants program using data from its \$1 billion ecology network. *Science*. <https://www.science.org/content/article/researchers-protest-end-nsf-grants-program-using-data-its-1-billion-ecology-network>
- Meyer, B. D., Mok, W. K., & Sullivan, J. X. (2015). Household surveys in crisis. *Journal of Economic Perspectives*, 29(4), 199-226.
- Parashar, M. (2024). Enabling Responsible Artificial Intelligence Research and Development through the Democratization of Advanced Cyberinfrastructure. *Harvard Data Science Review*.
- Parashar, M. (2025). Everywhere and Nowhere: Envisioning a Computing Continuum for Science. *Computing in Science & Engineering*, 27(1), 51-56.
- Parashar, M., & Altintas, I. (2023). Toward Democratizing Access to Science Data: Introducing the National Data Platform. 2023 IEEE 19th International Conference on e-Science (e-Science),
- Potok, N. (2022). Show US the Data. *Harvard Data Science Review*, 4(2).
- Rabbitt, M. P., Hales, L. J., Burke, M. P., & Coleman-Jensen, A. (2023). *Household food security in the United States in 2022*.
- Shuja, J., Alanazi, E., Alasmay, W., & Alashaikh, A. (2021). COVID-19 open source data sets: a comprehensive survey. *Applied Intelligence*, 51(3), 1296-1325.
- Skelley, G. (2025). Why it matters that Trump is deleting government data. In *ABC News*.

Statistics, B. o. J. (2024). *Survey of Inmates in Local Jails (SILJ), 2024-2025*.
(<https://bjs.ojp.gov/survey-inmates-local-jails-silj-2024-2025>)

Szalay, A. S., Kunszt, P., Thakar, A., Gray, J., & Slutz, D. (1999). The sloan digital sky survey and its archive. *arXiv preprint astro-ph/9912382*.

US Department of Agriculture. (2025). *USDA terminates redundant food insecurity survey*
<https://www.usda.gov/about-usda/news/press-releases/2025/09/20/usda-terminates-redundant-food-insecurity-survey>

Appendix

A.1. Comparison with alternative bibliographic sources

The Dimensions-derived dataset was validated against two additional bibliographic sources, Scopus and OpenAlex, using six datasets: ARMS, Census of Agriculture, Food Access Research Atlas, FoodAPS, HFSSM, and RUCC(L. Chenarides et al., 2025). Each source produced a distinct corpus of publications referencing these datasets. Scopus identified the largest number of DOIs for health- and economics-related datasets, while OpenAlex captured a broader range of publication types, including preprints and open-access journals. Dimensions provided more structured metadata linking datasets to publications, particularly through topic tags and institutional affiliations. For Scopus and OpenAlex, we constructed a defined search corpus to identify dataset mentions because neither platform supports full-text search. Across sources, overlap was limited—fewer than 10% of DOIs appeared in all three databases—reflecting differences in journal coverage, indexing scope, and search capabilities. These comparative results, combined with the availability of full-text access through the Dimensions API, informed our decision to use Dimensions as the primary data source for constructing the dataset.

A.2. USDA seed datasets and additional found datasets

USDA seed datasets

U.S. Department of Agriculture, National Agricultural Statistics Service. (n.d.). *Census of Agriculture (Ag Census)* [Data set]. Retrieved December 31, 2025, from <https://www.nass.usda.gov/AgCensus/> USDA

U.S. Department of Agriculture, National Agricultural Statistics Service. (n.d.). *Agricultural Resource Management Survey (ARMS)* [Data set]. Retrieved December 31, 2025, from https://www.nass.usda.gov/Surveys/Guide_to_NASS_Surveys/Ag_Resource_Management/ USDA

U.S. Department of Agriculture, Economic Research Service. (n.d.). *Food Access Research Atlas (FARA)* [Data set]. Retrieved December 31, 2025, from <https://www.ers.usda.gov/data-products/food-access-research-atlas> Economic Research Service

U.S. Department of Agriculture, Economic Research Service. (2025). *FoodAPS National Household Food Acquisition and Purchase Survey (FoodAPS)* [Data set]. Retrieved December 31, 2025, from <https://www.ers.usda.gov/data-products/foodaps-national-household-food-acquisition-and-purchase-survey> Economic Research Service

U.S. Department of Agriculture, Economic Research Service. (n.d.). *Household Food Security Survey Module (HFSSM)* [Survey instrument]. Retrieved December 31, 2025, from <https://www.ers.usda.gov/topics/food-nutrition-assistance/food-security-in-the-us/survey-tools#household> Economic Research Service

U.S. Department of Agriculture, Economic Research Service. (2025). *Food Security in the United States—Documentation (Current Population Survey Food Security Supplement)* [Data set documentation]. Retrieved December 31, 2025, from <https://www.ers.usda.gov/data-products/food-security-in-the-united-states/documentation> Economic Research Service

U.S. Department of Agriculture, Food and Nutrition Service. (n.d.). *Farm to School Census* [Data set]. Retrieved December 31, 2025, from <https://farmtoschoolcensus.fns.usda.gov/> USDA-FNS Farm to School Census

U.S. Department of Agriculture, Economic Research Service. (n.d.). *Using proprietary data (includes Circana/IRI InfoScan references)* [Webpage]. Retrieved December 31, 2025, from <https://www.ers.usda.gov/topics/food-markets-prices/food-prices-expenditures-and-establishments/using-proprietary-data> Economic Research Service

U.S. Department of Agriculture, National Agricultural Statistics Service. (n.d.). *Local Food Marketing Practices Survey* [Data set]. Retrieved December 31, 2025, from https://www.nass.usda.gov/Surveys/Guide_to_NASS_Surveys/Local_Food/index.php USDA

U.S. Department of Agriculture, Economic Research Service. (2024). *Quarterly Food-at-Home Price Database (1999–2010)* [Data set]. Retrieved December 31, 2025, from <https://www.ers.usda.gov/data-products/food-at-home-monthly-area-prices> Economic Research Service

U.S. Department of Agriculture, National Agricultural Statistics Service. (n.d.). *Tenure, Ownership, and Transition of Agricultural Land (TOTAL) Survey* [Data set]. Retrieved December 31, 2025, from https://www.nass.usda.gov/Surveys/Guide_to_NASS_Surveys/TOTAL/ USDA

U.S. Department of Agriculture, Economic Research Service. (2025). *Rural-Urban Continuum Codes (RUCC)* [Data set]. Retrieved December 31, 2025, from <https://www.ers.usda.gov/data-products/rural-urban-continuum-codes>

Additional found datasets

Table A.1. Additional datasets used in food security research, derived from LLM agent

Dataset Name	Agency / Organization Responsible
American Community Survey (ACS) Public Use Microdata Sample (PUMS)	US Census Bureau
American Community Survey 5-year estimates	US Census Bureau
American Community Survey (ACS) Single-year estimates (2018)	US Census Bureau
Integrated Public Use Microdata Series (IPUMS) CPS-FSS & ASEC	Collected by U.S. Census Bureau; harmonized/distributed by IPUMS (Univ. of Minnesota)
Canadian Community Health Survey (CCHS) Cycle 2.2 (2004)	Statistics Canada (joint initiative with Health Canada)
Food and Nutrient Database for Dietary Surveys (FNDDS)	USDA Agricultural Research Service (ARS)
Social Vulnerability Index (SVI) (CDC/ATSDR)	CDC / Agency for Toxic Substances and Disease Registry (ATSDR)
National Health Interview Survey (NHIS)	CDC / National Center for Health Statistics (NCHS)
American Time Use Survey (ATUS)	Sponsored by U.S. Bureau of Labor Statistics; conducted by U.S. Census Bureau
Behavioral Risk Factor Surveillance System (BRFSS) fruit & vegetable module	CDC (BRFSS) with state/territorial health departments
Food Environment Atlas	USDA/Economic Research Service (ERS)
Future of Families and Child Wellbeing Study (FFCWS) (formerly Fragile Families)	Princeton University (study home; access often via archives like ICPSR)
Household Pulse Survey (Phase 1)	U.S. Census Bureau (with federal partners)
University of Kentucky Center for Poverty Research National Welfare Data (UKCPR National Welfare Data)	University of Kentucky Center for Poverty Research (UKCPR)
American College Health Association - National College Health Assessment III (ACHANCHA III) survey	American College Health Association (ACHA)
Feeding America Map the Meal Gap (MMG)	Feeding America
Health and Retirement Study (HRS) — Health Care and Nutrition Study (HCNS)	Health and Retirement Study (University of Michigan ISR)
RAND HRS Longitudinal Dataset	Data collected by Health and Retirement Study (University of Michigan ISR); compiled by RAND Center for the Study of Aging

Note: Citations for each dataset are available below

Agency for Toxic Substances and Disease Registry. (n.d.). *CDC/ATSDR social vulnerability index (SVI)* [Data set]. Retrieved December 31, 2025, from <https://www.atsdr.cdc.gov/place-health/php/svi/index.html>

American College Health Association. (n.d.). *National College Health Assessment (NCHA): Survey reports* [Data set]. Retrieved December 31, 2025, from <https://www.acha.org/ncha/data-results/survey-results/all-ncha-survey-reports/>

Centers for Disease Control and Prevention. (n.d.). *National Health Interview Survey (NHIS)* [Data set]. Retrieved December 31, 2025, from <https://www.cdc.gov/nchs/nhis/index.html>

Centers for Disease Control and Prevention. (n.d.). *Data user's guide to the Behavioral Risk Factor Surveillance System (BRFSS) fruit and vegetable module* [Data set documentation]. Retrieved December 31, 2025, from <https://www.cdc.gov/nutrition/data-statistics/data-users-guide.html>

Feeding America. (n.d.). *Map the Meal Gap* [Data set]. Retrieved December 31, 2025, from <https://map.feedingamerica.org/>

Health Canada. (2006). *Canadian Community Health Survey (CCHS) Cycle 2.2, Nutrition (2004): A guide to accessing and interpreting the data* [Data set documentation]. Retrieved December 31, 2025, from <https://www.canada.ca/en/health-canada/services/food-nutrition/food-nutrition-surveillance/health-nutrition-surveys/canadian-community-health-survey-cchs/canadian-community-health-survey-cycle-2-2-nutrition-2004-guide-accessing-interpreting-data-health-canada-2006.html>

Health and Retirement Study, Institute for Social Research, University of Michigan. (n.d.). *2013 Health Care and Nutrition Study (HCNS)* [Data set]. Retrieved December 31, 2025, from <https://hrsdata.isr.umich.edu/data-products/2013-health-care-and-nutrition-study-hcns>

Minnesota Population Center. (n.d.). *IPUMS CPS: Current Population Survey (CPS) ASEC & Food Security Supplement (FSS)* [Data set]. Retrieved December 31, 2025, from <https://cps.ipums.org/cps/>

Princeton University. (n.d.). *Future of Families and Child Wellbeing Study (FFCWS)* [Data set]. Retrieved December 31, 2025, from <https://ffcws.princeton.edu/>

RAND Corporation. (n.d.). *RAND HRS longitudinal file (randhrs1992_2018v2)* [Data set]. Retrieved December 31, 2025, from <https://www.rand.org/health/surveys/hrs.html>

U.S. Bureau of Labor Statistics. (n.d.). *American Time Use Survey (ATUS)* [Data set]. Retrieved December 31, 2025, from <https://www.bls.gov/tus/>

U.S. Census Bureau. (n.d.). *American Community Survey (ACS) Public Use Microdata Sample (PUMS)* [Data set]. Retrieved December 31, 2025, from <https://www.census.gov/programs-surveys/acs/microdata.html>

U.S. Census Bureau. (n.d.). *American Community Survey (ACS) 5-year estimates (API/data sets)* [Data set]. Retrieved December 31, 2025, from <https://www.census.gov/data/developers/data-sets/acs-5year.html>

U.S. Census Bureau. (n.d.). *American Community Survey (ACS) 1-year estimates (2018)* [Data set]. Retrieved December 31, 2025, from <https://www.census.gov/newsroom/press-kits/2019/acs-1year.html>

U.S. Census Bureau. (n.d.). *Household Pulse Survey (Phase 1)* [Data set]. Retrieved December 31, 2025, from <https://www.census.gov/programs-surveys/household-pulse-survey.html>

U.S. Department of Agriculture, Agricultural Research Service. (n.d.). *Food and Nutrient Database for Dietary Studies (FNDDS)* [Data set]. Retrieved December 31, 2025, from <https://www.ars.usda.gov/northeast-area/beltsville-md-bhnrc/beltsville-human-nutrition-research-center/food-surveys-research-group/docs/fndds/>

U.S. Department of Agriculture, Economic Research Service. (n.d.). *Food Environment Atlas* [Data set]. Retrieved December 31, 2025, from <https://www.ers.usda.gov/data-products/food-environment-atlas>

University of Kentucky Center for Poverty Research. (n.d.). *UKCPR resources: National Welfare Data* [Data set]. Retrieved December 31, 2025, from <https://ukcpr.uky.edu/resources>

A.3. Identifying dataset joins

The LLM also described how datasets are combined within individual studies. The LLM-based joining agent scanned each full text for passages that described linking, merging, or otherwise integrating two or more datasets, such as through fusion, spatial joins, synthetic control, or key-based merges. For each identified case, it recorded the dataset pair, described how the datasets were combined, and categorized the type of linkage.

The resulting dataset-joins file contains one record per detected join (there may be multiple dataset joins per publication), with fields that summarize the linkage method, the reported join keys, and a short description of the integration drawn from the publication text. In total, this step identified 1,608 unique dataset pairings across 535 publications.

Documenting dataset joins is useful for understanding how data are used in applied research. The join patterns provide concrete examples of how researchers integrate data sources when studying food security, including which datasets are most frequently linked and which integration methods are used in practice. The file itself shows which dataset combinations are already common and provides context about the join, including reported integration challenges identified by the LLM-agent. Understanding data integration in practice is increasingly important because answering complex questions about food systems often requires combining multiple datasets within a single analysis (Lauren Chenarides et al., 2025). For those studying science and innovation, these patterns demonstrate how existing data systems are linked in practice and where established integration approaches could help researchers transition to alternative datasets when long-standing sources are reduced or retired.

A.4. Manual validation

Manual validation focused on the selection of newly discovered datasets. After the LLM extraction step (step 4), we obtained 77 candidate non-seed datasets. The LLM calculated a weighted score for each candidate that combined the number of publications in which it appeared and the LLM-assigned dataset use score, then sorted the list in descending order.

The candidates were manually reviewed in the scored order to determine whether each label referred to an actual dataset rather than a questionnaire, survey module, or generic phrase. For candidates judged to be real datasets, the topical relevance of their usage to food security was checked and their historical use in the corpus was examined. If multiple labels that referred to the same underlying dataset or to specific waves of a larger data product existed, they were consolidated. For example, “Health Care and Nutrition Study (HCNS)” and “RAND HRS longitudinal dataset (randhrs1992_2018v2)” were treated as the same dataset, since the latter label refers to a particular release of the broader RAND HRS series.

This manual review reduced the 77 candidates to a final set of 12 additional datasets that were retained for Stage Two of the workflow and incorporated into the data-usage descriptor tables.

A.5. Data Documentation

The workflow described in this study collectively produced a set of structured outputs, available in a GitHub repository: <https://github.com/national-data-platform/food-security-usage-descriptors>. The accompanying documentation describes field definitions for the datasets used in the research.

File A: USDA Publication-Dataset Pairs

fileA_usda_publication_dataset_pairs.csv is the primary data file generated through Step 3 of the workflow. Each record represents a verified instance where a dataset is mentioned within a publication.

Publications mentioning datasets were first detected using a string search approach with the Dimensions API, which identifies dataset references within the full text. These initial matches were then validated through the LLM-assisted workflow. Validation fields indicate whether the dataset was simply mentioned or used for analysis, along with the supporting text excerpt and model confidence scores.

All variables use standardized naming conventions and controlled vocabularies. Dataset names are harmonized to match canonical titles (for example, “Agricultural Resource Management Survey” instead of “ARMS”). Text fields preserve the original context to facilitate reproducibility and secondary analyses.

Table A.2 summarizes the file structure and variable definitions.

Table A.2.

Field name	Type	Description
publication_id	String	Unique identifier assigned to each publication within the Dimensions corpus.
publication_title	String	Full title of the publication.
publication_doi	String	Digital Object Identifier (DOI) for the publication.
validated_dataset_id	String	Internal identifier for the validated dataset entity.
validated_dataset_name	String	Canonical name of the dataset referenced in the publication.
confidence_score_mention	Float	Confidence score for dataset mentioned within the publication (0–10 scale).
confidence_score_use	Float	Confidence score that the dataset was used in the study’s analysis rather than only mentioned (0–10 scale).

File B: Publication-Dataset Pairs Using Alternative Datasets

fileB_additional_datasets_publication_pairs.csv documents newly identified datasets that were not included in the initial list of USDA datasets but were detected through the LLM. Each record represents a verified instance where a previously unlisted dataset was mentioned or used

within a peer-reviewed publication. The file extends the USDA-focused dataset inventory by capturing additional data sources that appear in the same research domain.

Each record links a publication to a newly discovered dataset, along with metadata describing the dataset’s name, context of use, and LLM-generated confidence scores for both identification and substantive usage. The file also provides contextual text excerpts, short dataset descriptions, and classification fields that indicate the dataset’s domain and source organization.

All dataset names are harmonized where possible to canonical titles, and metadata fields preserve the original text context to support reproducibility and follow-up verification.

Table A.3 summarizes the file structure and variable definitions.

Table A.3.

Field name	Type	Description
publication_id	String	Unique identifier assigned to each publication within the Dimensions corpus.
publication_title	String	Full title of the publication.
publication_doi	String	Digital Object Identifier (DOI) for the publication.
new_name	String	Canonicalized name of the newly discovered dataset.
new_confidence_score_mention	Float	Confidence score for dataset mentioned within the publication (0–10 scale).
new_confidence_score_use	Float	Confidence score that the dataset was used in the study’s analysis rather than only mentioned (0–10 scale).
new_description	String	Short description of the dataset as inferred from the publication context.
new_source	String	Organization or agency responsible for producing the dataset.
new_domain	String	Thematic or disciplinary domain of the newly identified dataset (for example, labor statistics, public health, or environmental monitoring). This reflects the dataset’s subject area rather than the publication’s research topic (for example, a food security study may use a dataset classified under the public health or labor domains).

File C: Dataset Join Table

fileC_dataset_joins.csv documents instances where multiple datasets are integrated within a single publication. Each record represents one observed dataset–dataset pairing (“join”) identified through the LLM. These records provide structured descriptions of how researchers

combine datasets, including the linkage methods, join keys, data-processing steps, and validation procedures.

The file includes technical and methodological details drawn directly from the publication text, providing a structured record of how data integration is described in applied research. It documents common challenges, quality control steps, and matching strategy used when linking distinct data sources.

All fields follow consistent naming conventions, with lists and nested attributes represented as JSON objects or arrays where appropriate. Text fields preserve verbatim context to support transparency and reproducibility.

Table A.4 summarizes the file structure and variable definitions.

Table A.4.

Field name	Type	Description
publication_id	String	Unique identifier assigned to each publication within the Dimensions corpus.
publication_title	String	Full title of the publication.
publication_doi	String	Digital Object Identifier (DOI) for the publication.
join_dataset1	String	Name of the first dataset in the join.
join_dataset2	String	Name of the second dataset in the join.
join_join_type	String	General category of data integration (for example, integration, linkage, or merge).
join_confidence_score	Float	Model-generated confidence score (0–10 scale) for correct identification of a dataset join.
join_methodology	String	Description of the procedure used to integrate the datasets, as described in the publication.
join_join_keys	Array	List of key variables used to link datasets (for example, tract_geoid, firm_id, year).
join_context	String	Narrative summary of how the dataset join contributes to the study’s objectives.

Supporting Files

Supporting File 1: Summary Table

summary_by_dataset_year.csv is a summary table that aggregates counts of publications by dataset and year. Each record includes the total number of unique DOIs referencing a given dataset within that year.

Supporting File 2: Data Dyads

data_dyads.csv includes the full list of dyads used for this search.

Supporting File 3: Metadata File

data_dictionary.csv is a metadata file that accompanies the dataset and provides definitions, data types, and valid values for all fields.

Supporting File 4: USDA Seed Datasets

usda_seed_datasets.csv is the list of the 11 USDA seed datasets (Stage 1).

Supporting File 5: Additional Datasets Discovered

additional_discovered_datasets.csv is the list of the 18 additional datasets (Stage 2).