Methodology for Comparing Citation Database Coverage of Dataset Usage

Findings

2025-06-18

Table of contents

Download PDF Version

Report Summary

What Is the Issue?

Federal datasets play an important role in supporting research across a range of disciplines. Measuring how these datasets are used can help evaluate their impact and inform future data investments. Agencies like the US Department of Agriculture (USDA) track how their datasets are referenced in research papers and disseminate data usage statistics through platforms like *Democratizing Data's* Food and Agricultural Research Data Usage Dashboard and NASS's 5 W's Data Usage Dashboard. These tools rely on identifying *dataset mentions*[1] in published research to develop usage statistics. Beyond reporting usage statistics, this type of analysis can also provide information about the research topics where federal datasets are applied. Understanding where federal datasets are applied helps characterize their disciplinary reach, including use in areas such as food security, nutrition, and climate, which are inherently multidisciplinary. This informs future work on identifying alternative datasets that researchers use to study similar questions across fields.

The process of identifying dataset mentions in academic research output has two requirements. First, citation databases provide structured access to large volumes of publication metadata, including titles, abstracts, authors, affiliations, and sometimes full-text content. Second, tracking dataset usage requires developing methods that scan publication text for dataset mentions. It is feasible to systematically identify where specific datasets are referenced across a broad set of research outputs by applying machine-learning algorithms to publication corpora collected from citation databases, allowing for scalable search and retrieval of relevant publications where datasets are mentioned. The accuracy of dataset tracking depends on the scope of research output we can access and analyze. However, different databases curate content (i.e., research output) in different ways - some focus on peer-reviewed journals while others include preprints and technical reports - and dataset tracking requires reliable citation data from citation databases.

This report presents a methodology for identifying dataset mentions in research publications across various citation databases. In doing so, we compare publication, journal, and topic coverage across Scopus, OpenAlex, and Dimensions [forthcoming] as primary sources. The purpose is to establish a consistent set of statistics for comparing results and evaluating differences in dataset tracking across citation databases. This allows for insights into how publication scope and indexing strategies influence dataset usage statistics.

[1] A dataset mention refers to an instance in which a specific dataset is referenced, cited, or named within a research publication. This can occur in various parts of the text, such as the abstract, methods, data section, footnotes, or references, and typically indicates that the dataset was used, analyzed, or discussed in the study.

How Was the Study Conducted?

Three citation databases are compared: Elsevier's Scopus, OurResearch's OpenAlex, and Digital Science's Dimensions.ai.

1. **Scopus** charges for access to its citation database. It indexes peer-reviewed, including journal articles, conference papers, and books, and provides metadata on authorship, institutional affiliation, funding sources, and citations. For this study, Scopus was used to identify dataset mentions through a two-step process: first, Elsevier executed queries against the full-text ScienceDirect corpus and reference lists within Scopus; second, publications likely to mention USDA datasets were filtered based on keyword matching and machine learning models.

2. **OpenAlex**, an open-source platform, offers free metadata access. It covers both traditional academic publications and other research outputs like preprints and technical reports. In this study, we used two approaches to identify dataset mentions in OpenAlex: a full-text search, which scans publication metadata fields such as titles and abstracts for references to USDA datasets,[2] and a seed corpus search, which starts with a targeted set of publications based on journal, author, and topic criteria, then downloads the full text of each paper to identify mentions of USDA datasets.[3]

3. **Dimensions**, developed by Digital Science, is a citation database that combines free and subscription-based access. It indexes a range of research outputs, including journal articles, books, clinical trials, patents, datasets, and policy documents. Dimensions also links publications to grant and funding information. For this study, publications in Dimensions that reference USDA datasets were identified by constructing structured queries in Dimensions' Domain Specific Language (DSL) that combined dataset aliases with institutional affiliation terms. These were executed via the `dimcli` API to return English-language articles from 2017–2023 with at least one U.S.-affiliated author. To maintain consistency with the criteria applied to Scopus and OpenAlex, the study focuses only on publications classified as journal articles.

To compare how these databases track dataset usage, we focus on six USDA datasets commonly used in agricultural, economic, and food policy research:

1. Agricultural Resource Management Survey (ARMS)
2. Census of Agriculture (Ag Census)
3. Rural-Urban Continuum Code (RUCC)
4. Food Access Research Atlas (FARA)
5. Food Acquisition and Purchase Survey (FoodAPS)

[2] Full-text search in OpenAlex refers to querying the entire database for textual mentions of dataset names within titles, abstracts, and other fields.

[3] The seed corpus search involves selecting a targeted set of publications based on journal, author, and topic filters. Full-text PDFs are downloaded and analyzed to identify mentions of USDA datasets not captured through metadata alone.

6. Household Food Security Survey Module (HHFSS)

These datasets were selected for their policy relevance, known usage frequency, and disciplinary breadth. We developed seed corpora for each dataset to identify relevant publications, then used those corpora to evaluate database coverage, topical scope, and metadata consistency.

What Did the Study Find?

Accurate tracking of dataset mentions relies heavily on how publications are indexed across citation databases. For two citation databases – Scopus and OpenAlex – carefully constructed seed corpora were needed to track dataset mentions.

Preview of Results from Database Comparison:

1. Across databases, there is limited publication overlap between citation databases. For example:

- Less than 10% of DOIs typically appear in both Scopus and OpenAlex in any combination.
- 51.8% of Food Access Research Atlas DOIs appear only in Scopus.
- 60.9% of Household Food Security Survey Module DOIs appear only in Scopus.
- 78.5% of ARMS DOIs appear only in OpenAlex Full Text.

2. Journal coverage by source (Scopus or OpenAlex) varies significantly by dataset:

- Scopus recovers the most publications MORE HERE.
- OpenAlex "Full Text" recovers the most publications MORE HERE.
- OpenAlex "Seed Search" identifies the most publications MORE HERE.

3. Topical coverage reflects the varied policy and disciplinary relevance of each dataset:

- ARMS: Research citing this dataset emphasizes agricultural management, accounting, and environmental topics.
- The Census of Agriculture: Research mentioning this dataset has a wide breadth, spanning accounting and environmental applications.
- Food Access Research Atlas: Publications focus on food security, public health, and urban planning.
- The Food Acquisition and Purchase Survey: This dataset is mentioned in studies of consumer behavior, nutrition economics, and household spending.
- The Household Food Security Survey Module: Research mentioning this dataset frequently cites topics such as food insecurity, poverty, and social policy evaluation.
- The Rural-Urban Continuum Code: Research citing this dataset includes rural classification, regional planning, and spatial analysis.

Key Takeaway: These patterns suggest that relying on a single citation database may undercount dataset usage, and may also obscure variation in the types of research topics being conducted with each dataset.

How to Use This Report

The report is preliminary in nature. It provides an initial approach to characterizing dataset mentions about food and agriculture research datasets in research papers reported in various databases, specifically Scopus, OpenAlex, and Dimensions. It includes procedures for:

- Identifying publication coverage across citation databases
- Cross-referencing publications between datasets
- Analyzing research themes and institutional representation

The methodology produced these reusable components:

- Code repository for data cleaning and standardization
- Data schemas by citation database
- Standardized institution tables using IPEDS identifiers

The methods described can be applied to evaluate other citation databases such as Web of Science, Crossref, and Microsoft Academic, to name a few.

1 Project Background

Tracking how federal datasets are used in academic research has been a priority for agencies such as the U.S. Department of Agriculture (USDA). *Democratizing Data's* Food and Agricultural Research Data Usage Dashboard was created to support this effort by reporting on dataset usage through citation analysis. The platform was developed to ingest publication metadata from Scopus, a proprietary citation database, to identify and count publications that reference USDA datasets. Scopus offers reliable metadata and a structured indexing system, bu it is costly to access and does not fully align with goals around open science and public transparency.

As interest in open-access infrastructure has grown, OpenAlex, a free and open-source citation database developed by OurResearch, has emerged as a potential alternative. OpenAlex claims broad coverage of research outputs, including journal articles, preprints, conference proceedings, and reports. Replacing Scopus with OpenAlex could lower operational costs for federal agencies and align with broader efforts to promote open data ecosystems. However, transitioning platforms raises important questions about data reliability, coverage completeness, and potential trade-offs in representation.

To support an informed decision about this transition, a systematic comparison was conducted across three citation databases—Scopus, OpenAlex, and Dimensions—to assess their relative strengths and weaknesses for tracking dataset mentions in agricultural and food systems research. Dimensions, a third database developed by Digital Science, offers a hybrid model combining free and subscription-based access and was included to provide a broader benchmark across commercial and open platforms.

Initial comparisons between Scopus and OpenAlex revealed unexpected differences in coverage, with notable gaps in publication indexing and metadata quality. These patterns suggest that simply substituting one citation source for another could lead to incomplete or biased tracking of dataset usage, potentially affecting public reporting and research visibility. This project responds to those concerns by developing a structured, reproducible methodology for evaluating database coverage across multiple dimensions: publication metadata, journal inclusion, dataset topic area, institutional affiliation, and authorship.

1.1 Project Objective

This report presents a method for tracking how six key USDA datasets (Table 1) are cited in research using Scopus, OpenAlex, and Dimensions. It identifies where each dataset appears, which topics they are used in, which authors and institutions are most active, and how these patterns vary depending on the citation database. The findings support more accurate measurement of dataset use and help guide future data preservation and investment decisions.

1.2 Specific Aims

1. **Evaluate differences in publication coverage across citation databases.** Measure the extent to which Scopus, OpenAlex, and Dimensions capture research publications that reference USDA datasets. Identify how publication inclusion varies across platforms.
2. **Compare journal indexing and scope.** Compare the journals indexed by each database and examine how differences in journal coverage influence visibility of dataset-linked research.
3. **Analyze topic coverage.** Examine the research areas where USDA datasets are mentioned. Identify patterns in topic classification and assess how different citation databases support subject-level tracking of dataset usage.
4. **Evaluate author representation.** Compare how author names are recorded across platforms, including the completeness of author metadata and potential implications for attribution and visibility.
5. **Examine institutional representation.** Evaluate how each platform captures and standardizes institutional affiliations. Pay particular attention to differences in coverage for Minority-Serving Institutions (MSIs), land-grant universities, and other public or underrepresented institutions.
6. **Develop a reproducible methodology for cross-platform comparison.** Create a generalizable workflow for comparing citation databases, including steps for record linkage, deduplication, author and institution standardization, and identification of dataset mentions.

The methodology described in this report provides a systematic approach for comparing publication coverage where federal datasets are mentioned across citation databases. The scope of work includes comparing publication coverage across Scopus, OpenAlex, and Dimensions. For more information on each citation database, refer to this Appendix. These methods can be applied to other citation databases as alternatives to current data sources.

2 Data Collection

The core objective of this study is to evaluate publication coverage across citation databases, focusing on how well Scopus, OpenAlex, and Dimensions index research relevant to food and agricultural research. A targeted strategy was used to identify publications referencing USDA datasets, aligning with federal agency efforts to monitor and report on dataset usage. This approach enables a consistent entry point for comparison across platforms while also providing insight into the topics where federal datasets are applied and the use of complementary or alternative data sources.

To support this analysis, a structured inventory of USDA data assets was developed, drawing from records produced by the Economic Research Service (ERS) and the National Agricultural Statistics Service (NASS). From this broader inventory, six datasets were selected for detailed comparison based on known usage, policy relevance, and disciplinary breadth: the Census of Agriculture, Agricultural Resource Management Survey (ARMS), Food Acquisition and Purchase Survey (FoodAPS), Food Access Research Atlas (FARA), Rural-Urban Continuum Code (RUCC), and the Household Food Security Survey Module (HFSSM). The set of data assets, their producing agencies, and descriptions are presented in Table 1.

Table 1: List of USDA Data Assets

Dataset Name	Produced By	Description
Census of Agriculture	NASS	Conducted every five years, it provides comprehensive data on U.S. farms, ranches, and producers.
Agricultural Resource Management Survey (ARMS)	ERS	A USDA survey on farm financials, production practices, and resource use.
Food Acquisition and Purchase Survey (FoodAPS)	ERS	A nationally representative survey tracking U.S. household food purchases and acquisitions.
Food Access Research Atlas (FARA)	ERS	A USDA tool mapping food access based on store locations and socioeconomic data.
Rural-Urban Continuum Code (RUCC)	ERS	A classification system distinguishing U.S. counties by rural and urban characteristics.
Household Food Security Survey Module	ERS	A USDA survey module used to assess food insecurity levels in households.

Researchers reference datasets in inconsistent ways—using acronyms, abbreviations, alternate spellings, or related URLs. To capture these variations, we created a structured list of dataset–alias pairs, called *dyads*. This Appendix provides the full list of dyads used to search for mentions of each USDA dataset across Scopus, OpenAlex, and Dimensions. This list ensures consistent and comprehensive identification of dataset mentions in research publications.

Using these dyads, we applied tailored search strategies across each citation database to identify

relevant publications for all six datasets. These included a seed search in Scopus, a full-text metadata search in OpenAlex, a seed corpus approach in OpenAlex based on targeted filtering of journals, authors, and topics followed by full-text analysis, and a full-text search in Dimensions. Each search strategy is described in detail in the following sections.

2.1 Scopus Approach

The first citation database used is Scopus, a publication catalog managed by Elsevier. Ideally, direct Scopus API access would have been used to query full publication text for mentions of USDA datasets. However, the project did not have access to the Scopus API. Only Elsevier, serving as a project partner, was able to execute queries within the Scopus environment. Consequently, the dataset mention search relied on outputs provided by Elsevier rather than independent querying.

Because of these constraints, a seed corpus approach was applied. First, Elsevier matched the names and aliases of all USDA datasets against full-text records available through ScienceDirect and reference sections of Scopus publications published between 2017 and 2023. This initial step identified journals, authors, and topics most likely to mention USDA datasets. A targeted search corpus was then constructed, narrowing the scope to approximately 1.45 million publications. These included various document types—articles, reviews, short surveys, notes, conference papers, chapters, books, editorials, letters, data papers, errata, and tombstones. For the purposes of this comparative report, only articles are considered.

Several methods were used to identify mentions of USDA datasets in Scopus publications. First, a reference search was conducted, using exact-text matching across publication reference lists to capture formal citations of datasets. Second, full-text searches were performed using machine learning models applied to publication bodies, identifying less formal mentions of datasets. Third, machine learning routines developed through the 2021 Kaggle competition were applied to the full-text corpus to improve detection of dataset mentions, including instances where references were indirect or less structured. Details about the three machine learning models used are available here.

Because direct access to full publication text was not available, Elsevier shared only the extracted snippets and limited metadata. Manual validation, aided by the use of keyword flags (e.g., "USDA," "NASS"), confirmed whether identified mentions accurately referred to the targeted datasets. To manage validation costs, only publications with at least one U.S.-based author were reviewed.

Full documentation of the Scopus search routine, including query construction and extraction procedures, is available at the project's report website.

2.2 OpenAlex Approach

The second citation database used is OpenAlex, an open catalog of scholarly publications that provides public acess to metadata and, when available, full-text content for open-access publications via its API. Unlike Scopus, which provides controls access to licensed content, OpenAlex indexes only open-access publications or those for which open metadata has been made available by publishers.

Two methods were used to identify USDA dataset mentions in OpenAlex: a full-text search and a seed corpus approach. Both Both methods focused on peer-reviewed journal articles published between 2017 and 2023 and restricted the dataset to final published versions, excluding preprints and earlier drafts to avoid duplication across versions.

2.2.1 Method 1: Full-Text Search

This method relied on querying OpenAlex's full-text search index using combinations of dataset aliases (e.g., alternate names, acronyms) and institutional flag terms (e.g., "USDA," "NASS"). The combination of dataset alias and flag terms ensured that retrieved publications made an explicit connection to the correct data source. A "true" dataset mention was recorded only when at least one alias and one flag term appeared in the same publication, increasing the precision of captured dataset mentions.[4]

Queries were implemented using the `pyalex` Python package[5], which manages API requests and enforces OpenAlex's usage rate limits. The search used the `search` and `filter` endpoints, targeting English-language, open-access articles or reviews published after 2017. Results were returned in JSON format based on the OpenAlex Work object schema, including fields for publication metadata, authorship, journal, concepts, citations, and open access status. Each record included metadata fields such as:

- `display_name` (publication title)
- `authorships` (authors and affiliations)
- `host_venue.display_name` (journal)
- `doi` (digital object identifier)
- `concepts` (topics)

[4] This procedure increased the likelihood of capturing genuine dataset references rather than incidental matches to individual words. Initial drafts of the query incorrectly included terms like "NASS" and "USDA" in the alias list. This was corrected to ensure that aliases strictly referred to dataset names, and flag terms referred to organizations.

[5] `Pyalex` is an open-source library designed to facilitate interaction with the OpenAlex API; see https://help.openalex.org/hc/en-us/articles/27086501974551-Projects-Using-OpenAlex for more information. The package manages request formatting and automates compliance with OpenAlex's "polite pool" rate limits, which restrict the number of requests per minute and impose backoff delays. Pyalex introduced automatic pauses between requests, with a default `retry_backoff_factor` of 100 milliseconds, to ensure stable and continuous retrieval. This setup enabled systematic querying while adhering to OpenAlex's usage policies.

- `cited_by_count` (citation counts)
- `type` (publication type, e.g., "article")
- `publication_year` (year article was publish)
- `language` (language, English only)
- `is_oa` (open access)

The code used to implement this querying and filtering process is publicly available here **ADD FILES**.

2.2.1.1 Limitations of Full-Text Search Method

Although the OpenAlex API provides access to full-text search, limitations in content ingestion affect result completeness. OpenAlex receives publication text through two primary ingestion methods: PDF extraction and n-grams delivery.

In the PDF ingestion method, OpenAlex extracts text directly from the article PDF. However, the references section is not included in the searchable text. References are processed separately to create citation pointers between scholarly works, meaning that mentions of datasets appearing only in bibliographies are not discoverable through full-text search.

In the n-grams ingestion method, OpenAlex does not receive the full article text. Instead, it receives a set of extracted word sequences (n-grams) from the publisher or author. These n-grams represent fragments of text—typically short sequences of one, two, or three words—which are not guaranteed to preserve full continuous phrases. As a result, complete dataset names may be broken apart or omitted, reducing the likelihood that search queries match the intended aliases.

These ingestion and indexing limitations affect the completeness of results when relying solely on OpenAlex full-text search. Mentions of USDA datasets that appear either exclusively in references or are fragmented within n-grams may be missed. To address these limitations, an alternative search method was developed based on constructing a filtered seed corpus of publications for local full-text analysis.

2.2.2 Method 2: Seed Corpus

To overcome the limitations of the full-text metadata search, a seed corpus approach was developed. This method created a filtered subset of publications for local full-text analysis, targeting likely mentions of USDA datasets.

Selection criteria for the seed corpus included:

- English-language publications
- Works published between 2017-2023
- Publication Type = articles

- Open-access publications only

To focus the sample, we used results from the initial OpenAlex full-text search to identify the top 25 journals, authors, and topics most frequently associated with USDA dataset mentions. For each entity, we computed a *Full-Text Search Count*, which is the number of publications where USDA datasets were explicitly mentioned in the full text. This metric reflects how often each topic, journal, or author has appeared in USDA dataset–relevant research.

We then filtered the broader OpenAlex catalog to include all publications—regardless of whether they mentioned a dataset—linked to these top-ranked entities. This allowed us to build a more focused but expansive corpus for local text search. By narrowing to 25 entities per category, we prioritized relevance while managing scale. This process generated a structured set of JSON files containing publication metadata and links. The Python script used to flatten and process these files is provided in this Appendix. **ADD FILES**

Example: Census of Agriculture

To illustrate this process, consider the tables created for the Census of Agriculture dataset—Table 6 (top 25 topics), Table 7 (top 25 journals), and Table 8 (top 25 U.S.-affiliated authors). Each table contains two columns:

- **Full-Text Search Count**: Number of publications from the OpenAlex full-text search that mention the dataset and are linked to the given topic, journal, or author
- **Total Count**: Total number of publications in OpenAlex associated with that topic, journal, or author, regardless of dataset mention

The *Full-Text Search Count* helps us identify which entities are most directly associated with USDA dataset use. For instance, if a topic like "Impact of Food Insecurity on Health Outcomes" has 78 dataset-related publications. This count reflects how often USDA datasets were mentioned within the full text of publications associated with a particular entity. Meanwhile, the *OpenAlex Total Count* shows the broader publication volume for that topic—in this case, over 78,000—providing context on how prominent the topic is within the full OpenAlex database. In this sense, the Full-Text Search Count serves as a rough proxy for market penetration, or how frequently a dataset appears within a given research area relative to the total volume of publications.

The Full-Text Search Count reflects how often USDA datasets are explicitly mentioned within a specific research area, while the Total Count represents the overall volume of publications linked to that topic, journal, or author. The large gap between these counts was a key reason for developing the seed corpus approach: even within high-relevance entities, many publications may reference datasets in ways not captured by OpenAlex's full-text search.

By downloading and analyzing the full texts of all publications linked to the entities in the second column, we applied our own string-matching logic to detect mentions that OpenAlex's indexing may have missed, particularly in reference sections or when dataset names were

fragmented. This allowed us to validate and extend OpenAlex search results using a consistent and transparent local method.

This approach has several implications. It increases the relevance of the corpus by focusing on publications where USDA datasets are actively cited, rather than broadly associated with a topic. It also reduces processing demands by avoiding the need to download all potentially relevant PDFs. However, by prioritizing high-visibility entities from the initial search, the method may introduce selection bias and miss less frequently cited but still relevant work. The trade-off reflects a practical balance between analytical depth and operational feasibility.

For the Census of Agriculture, the resulting seed corpus included approximately 1.77 million unique publications. About 35% of full texts were successfully downloaded, yielding an estimated 625,000 documents for local analysis. Full-text searches on this subset improved detection of dataset mentions beyond what OpenAlex's native indexing allowed.

Despite the benefits, limitations remain. Full-text availability was constrained by broken or inaccessible links, and processing the corpus was computationally intensive. Future work may require distributed processing or more refined filters to improve efficiency.

The table below summarizes primary differences between the Full-Text Search and Seed Corpus methods. The Full-Text Search provides broader initial coverage, but it is limited by indexing constraints and lack of reference section access. The Seed Corpus narrows the search space but allows for deeper, locally controlled analysis of full-text content, including citations.

Table 2: Key Differences Between OpenAlex Full-Text Search and Seed Corpus

Feature / Criterion	Full-Text Search	Seed Corpus
Searchable Sample	OpenAlex API where `has_fulltext = true`	Curated list based on known users/sources
Source of text	Article body or word/phrase snippets where `fulltext_origin = n-grams`	Any part of publication conditional on available PDF download
Reference sections indexed?	No	Yes. Will include publications that reference datasets in citations.
Full text required? (`has_fulltext`)	Yes	Not required
Open access required? (`is_oa`)	No	Yes. Method requires downloading the full PDF version of the article.
Selection criteria	None imposed *a priori*	Journal/topic/author targeting
Resulting sample	Broad, but with limitations	Narrower, given the target search criteria

2.3 Dimensions

To identify publications mentioning USDA datasets, we used the Dimensions.ai API, following the same general methodology applied in Scopus and OpenAlex. We reused the same dataset aliases, institutional flag terms, and overall search criteria to ensure consistency across sources. The search covered scholarly publications from 2017 to 2023 and was restricted to works authored by at least one researcher affiliated with a U.S.-based institution.

Dimensions queries are written using a structured Domain Specific Language (DSL). We constructed boolean queries that combined multiple dataset aliases (e.g., "NASS Census of Agriculture", "USDA Census", "Agricultural Census") with institutional identifiers (e.g., "USDA", "NASS", "U.S. Department of Agriculture"). As with Scopus and OpenAlex, both a dataset alias and an institutional flag term were required to appear in each result. These terms were grouped using `OR` within each category and then combined with an `AND` across categories. For example:

> ("NASS Census of Agriculture" OR "Census of Agriculture" OR "USDA Census of Agriculture" OR "Agricultural Census" OR "USDA Census" OR "AG Census") AND (USDA OR "US Department of Agriculture" OR "United States Department of Agriculture" OR NASS OR "National Agricultural Statistics Service")

We implemented this process using the `dimcli` Python library, which provides a streamlined interface to the Dimensions.ai API and automates result pagination. A significant advantage of this approach is the capability of the Dimensions.ai platform to manage complex searches directly, resulting in precise results and reduced computational overhead. By executing these queries directly through the API, we avoided the technical complexity associated with downloading and locally processing large amounts of textual content. Moreover, the Dimensions.ai API results can be automatically structured into an analysis-ready DataFrame format. This simplified data structure greatly facilitated our subsequent validation, data integration, and analytical workflows.

To maintain methodological consistency with Scopus and OpenAlex, the following filters were applied to the search:

- English-language publications
- Works published between 2017-2023
- Document types: articles, chapters, proceedings, monographs, and preprints
- Author affiliations: Publications were filtered to include only those authored by researchers affiliated with at least one U.S.-based institution.

For comparability with the Scopus and OpenAlex samples, only publications classified as "articles" were retained for final analysis. This restriction reduces duplication across versions (e.g., preprints, proceedings) and reflects our focus on peer-reviewed scholarly output.

For each article, we retrieved metadata including title, authors, DOI, journal, abstract, publication date, citation counts, subject classifications, and links. These fields supported topic-level analysis, author and institution mapping, and validation of dataset mentions.

Using Dimensions.ai provided two main technical advantages. First, because the platform supports full-text query execution natively, we avoided the need to download or parse external files. Second, the API responses were easily converted into analysis-ready DataFrames, which simplified downstream validation and integration with other sources.

Overall, the Dimensions.ai approach aligned with our methods for Scopus and OpenAlex, enabling consistent identification of USDA dataset mentions across all three platforms.

2.4 Data Processing

To produce a consistent count of unique publications referencing each USDA dataset, records from three sources-Scopus, OpenAlex, and Dimensions-were consolidated, each of which identified publications through a different mechanism, described above.

For each source, publication-level metadata, including DOIs, journal titles, ISSNs (when available), and source-specific topic classifications was extracted. DOIs were standardized (e.g., removing URL prefixes, `https://doi.org/`) for consistent matching across sources. Duplicate DOIs within each source were removed. All DOIs compared in this report are associated with publications classified as document type = `article` and were published between 2017 and 2023.

3 Results

The aims described in Section 1.2 guide the development of a methodology for comparing citation databases, focusing on four areas:

1. **Publication tracking:** Comparing how each platform captures publications within indexed journals
2. **Journal coverage:** Determining which journals each platform indexes
3. **Topic scope:** Evaluating the research areas of publications that cite USDA datasets
4. **Author and institutional affiliation:** Determining how each platform records institutional information

Processed publication metadata was then merged across sources using the cleaned DOI-ISSN pairs as the common identifier. Each publication was tagged with binary indicators showing whether it appeared in Scopus, OpenAlex Full Text, OpenAlex Seed, or some combination

thereof. When metadata overlapped (such as journal titles or publication years), Scopus information was prioritized, when available, given its relatively higher metadata quality, followed by OpenAlex Full Text, OpenAlex Seed, and then Dimensions.[6]

This process ensured that each publication was counted once, even if it appeared in multiple sources. The final dataset includes a deduplicated set of DOIs, along with harmonized metadata and source indicators. The number of unique publications referencing each dataset is shown in Table 3.

Table 3: Unique Publications with Metadata across Sources

Dataset Name	Number of Unique Publications
ARMS	1,581
Census of Agriculture	5,835
Food Access Research Atlas	590
Food Acquisition and Purchase Survey	808
Household Food Security Survey Module	1,408
Rural-Urban Continuum Code	2,215

All code used to clean, deduplicate, and merge records is provided in the GitHub repository.

3.1 Publication Coverage

An objective of this report is to understand differences in publication coverage across Scopus, OpenAlex, and Dimensions. Specifically, this section asks: (1) how many and which publications referencing USDA datasets appear in each citation database, and (2) how many and which journals publishing these articles overlap between the two sources.

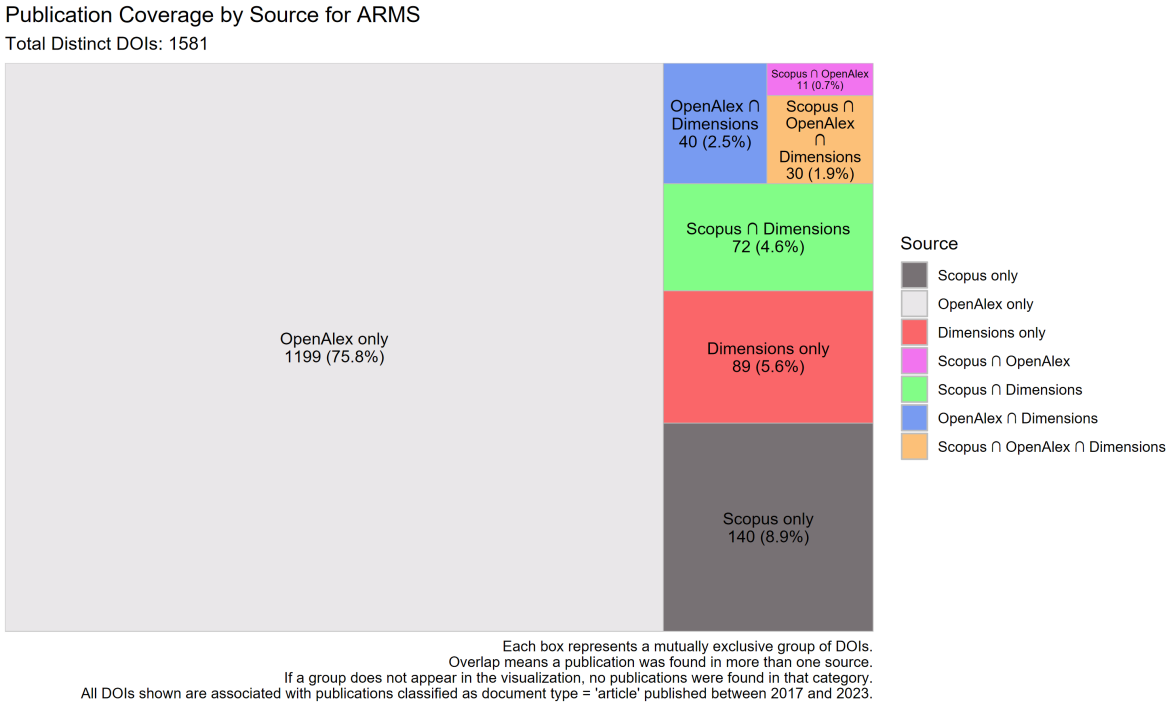
In addition, the analysis evaluates whether the different search strategies used in OpenAlex—the full-text metadata search versus the seed-corpus approach—yield substantially different sets of results.

For each of the six USDA datasets (Table 1) featured in this study, a treemap visualization summarizes publication coverage across the three citation databases. Each treemap groups publications into mutually exclusive categories based on their presence in one or more of the sources. The size of each box is proportional to the number of distinct DOIs in that group, providing a visual summary of relative coverage. For example, a large "Scopus only" segment indicates a high number of publications indexed exclusively in Scopus, while overlapping segments (e.g., "Scopus Dimensions") reflect shared coverage between platforms.

[6] In cases where a publication appeared in more than one source, manual and programmatic checks confirmed that metadata values, such as journal titles and publication years, were consistent across sources. No conflicting values were detected.

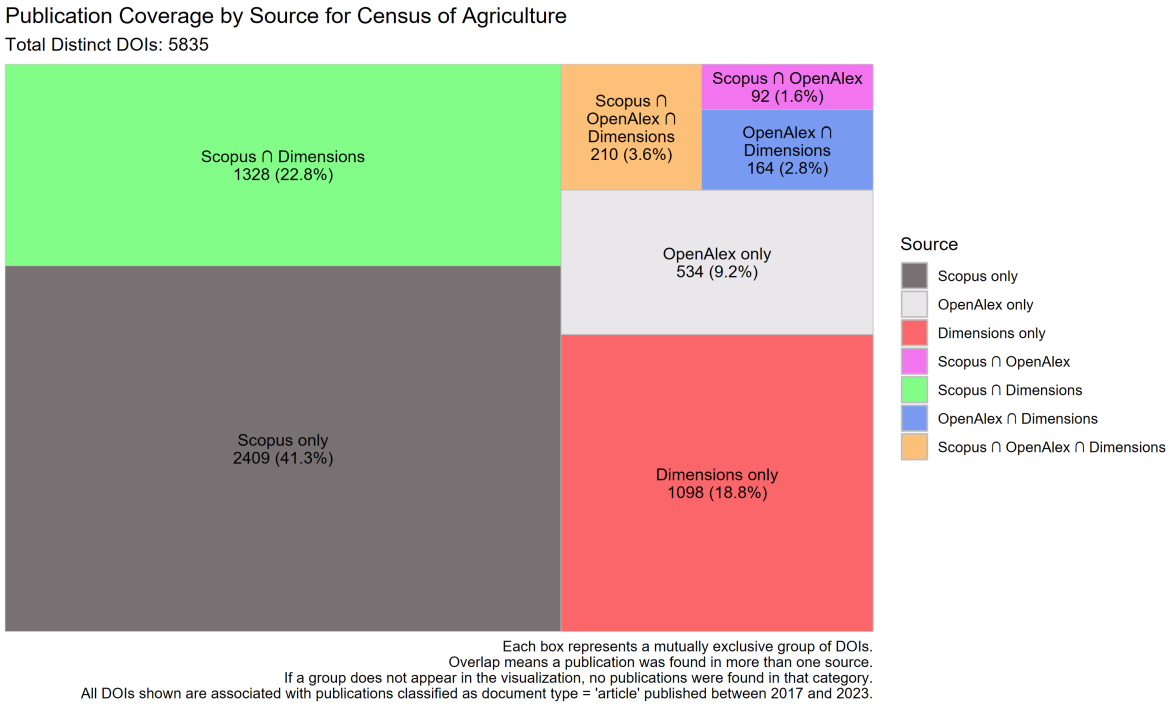
Agricultural Resource Management Survey (ARMS)

OpenAlex dominates coverage for ARMS-related publications, capturing nearly 76% of all distinct DOIs exclusively. In contrast, Scopus and Dimensions contribute relatively little: just 8.9% and 5.6% of DOIs appear exclusively in those sources, respectively. Overlaps are modest, with 2.5% of DOIs shared by OpenAlex and Dimensions, and only 1.9% captured by all three. This suggests OpenAlex's broader indexing of ARMS publications relative to the other databases.



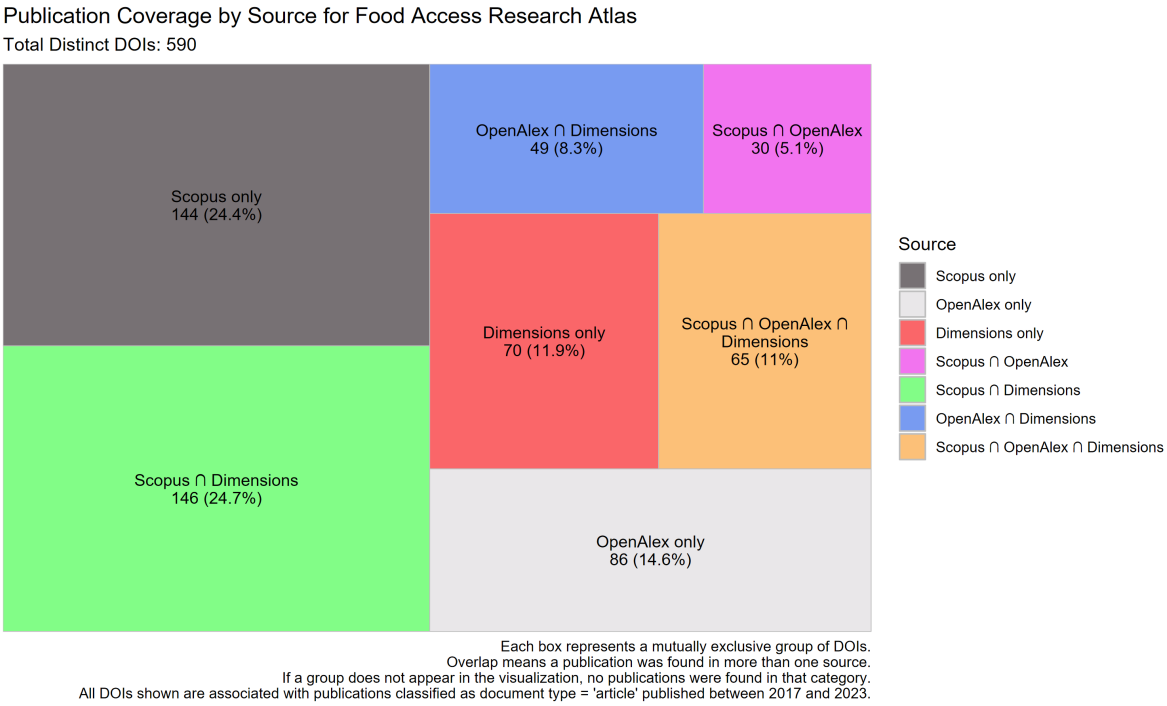
The Census of Agriculture

Scopus provides the broadest exclusive coverage for the Census of Agriculture, accounting for 41.3% of DOIs. Dimensions follows at 18.8%, while OpenAlex accounts for just 9.2% exclusively. The largest overlap is between Scopus and Dimensions (22.8%), with limited three-way overlap (3.6%). These results indicate that Scopus and Dimensions are the primary sources capturing publications referencing this dataset.



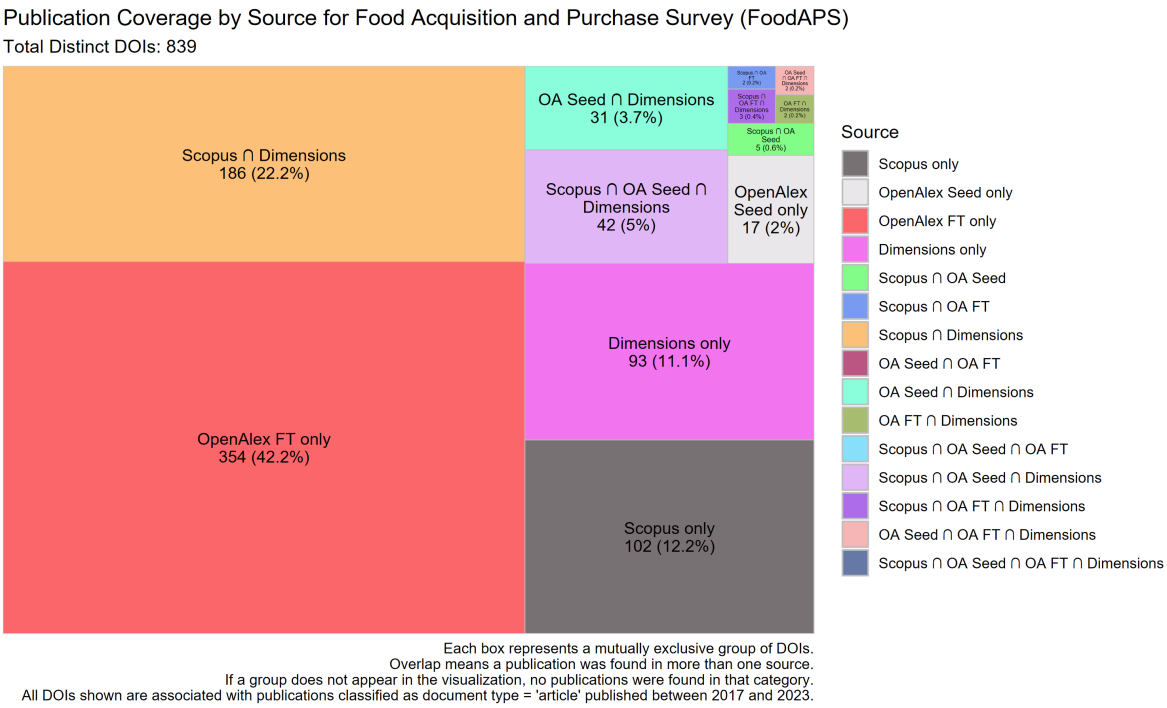
Food Access Research Atlas

Coverage for this dataset is more evenly distributed. Scopus and Scopus ∩ Dimensions each account for about 24%, while OpenAlex-only coverage is 14.6%, and Dimensions-only is 11.9%. Notably, 11% of DOIs appear in all three sources. This more balanced distribution suggests broader and more consistent indexing across platforms, without a single source dominating.



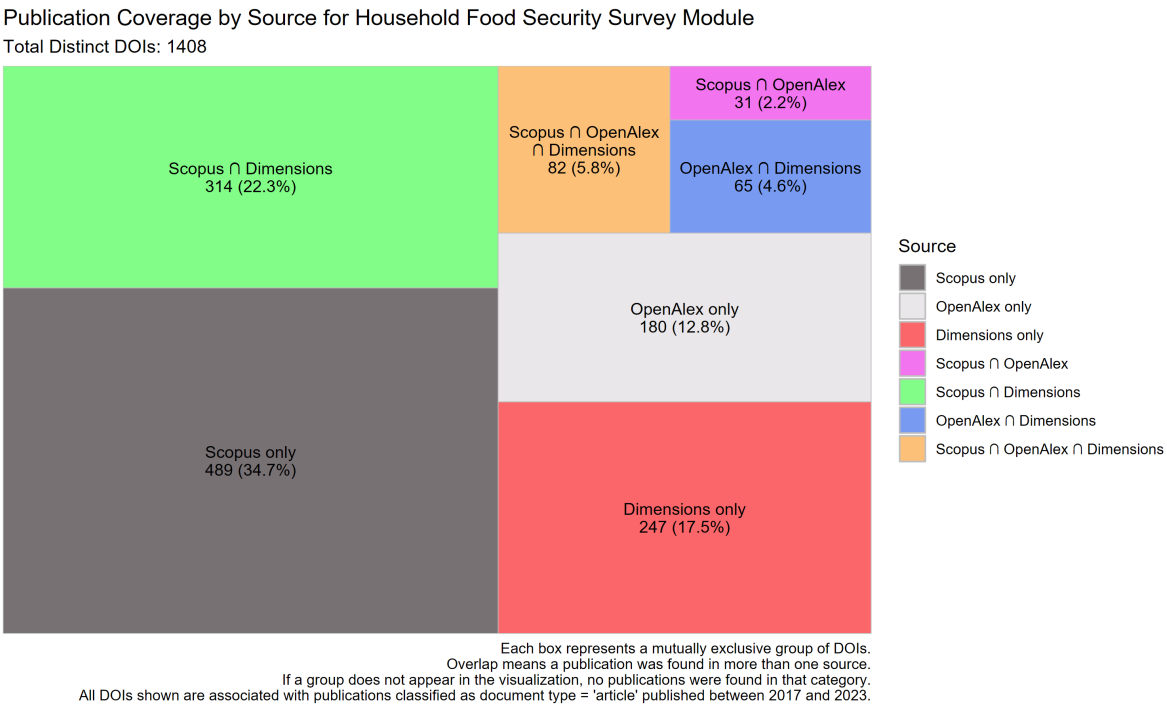
The Food Acquisition and Purchase Survey (FoodAPS)

OpenAlex again provides the widest exclusive coverage (46.7%), while Scopus and Scopus Dimensions each contribute 17.8%. Dimensions-only coverage is modest (7.7%), and 4.7% of DOIs are shared across all three. This indicates that OpenAlex is especially important for capturing FoodAPS-related work, but combined use of all three sources increases overall visibility.



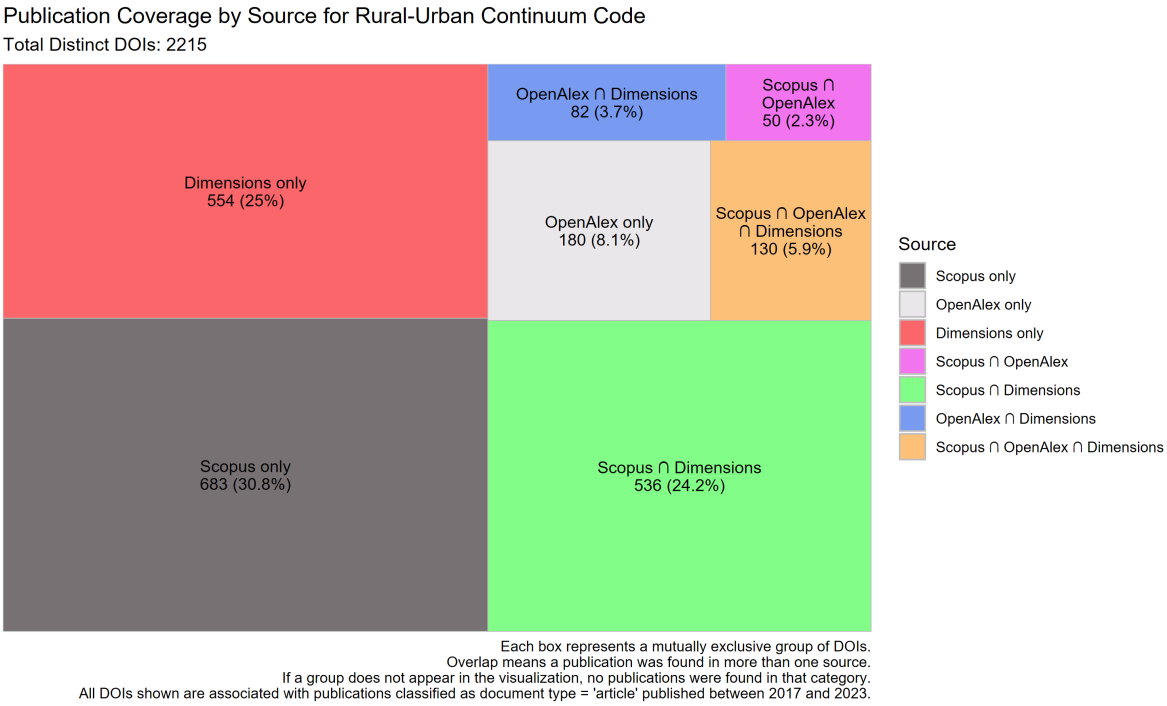
The Household Food Security Survey Module

Scopus has the highest exclusive coverage (34.7%), followed by Scopus ∩ Dimensions (22.3%) and Dimensions-only (17.5%). OpenAlex-only coverage is lower at 12.8%, and just 5.8% of DOIs are indexed by all three. This indicates stronger coverage for HFSSM-related publications in Scopus and Dimensions compared to OpenAlex.



Rural-Urban Continuum Code

Coverage is again led by Scopus (30.8%) and Dimensions (25%), with Scopus Dimensions contributing another 24.2%. OpenAlex-only coverage is relatively low at 8.1%, and only 5.9% of DOIs are shared across all three. This pattern is consistent with datasets where OpenAlex's coverage is more limited.



Synthesis of DOI Coverage by Source (Percent of Total DOIs)

Dataset	Total DOIs	Scopus only (%)	OpenAlex only (%)	Dimensions only (%)	Scopus ∩ OpenAlex (%)	Scopus ∩ Dimensions (%)	OpenAlex ∩ Dimensions (%)	All three (%)
ARMS	1581	8.9	75.8	5.6	0.7	4.6	2.5	1.9
Census of Agriculture	5835	41.3	9.2	18.8	1.6	22.8	2.8	3.6
Food Access Research Atlas	590	24.4	14.6	11.9	5.1	24.7	8.3	11.0
FoodAPS	808	17.8	46.7	7.7	1.7	17.8	3.6	4.7
HFSSM	1408	34.7	12.8	17.5	2.2	22.3	4.6	5.8
RUCC	2215	30.8	8.1	25.0	2.3	24.2	3.7	5.9

3.2 Journal Coverage

The previous section documented substantial variation in publication coverage across Scopus, OpenAlex, and Dimensions. One potential explanation for these differences is variation in journal indexing across sources. This section examines that possibility by looking at journal coverage, specifically, whether each citation database indexes the journals where USDA dataset-related publications appear.

For each dataset, the analysis identifies the top 40 journals (by DOI count) and determines which citation databases index them. Sankey diagrams illustrate the relationship between citation databases (left) and journals (right). Flows indicate coverage, with journals indexed in multiple sources connected to each. While only the top 40 journals are visualized, a complete list is available in the GitHub repository.

Agricultural Resource Management Survey (ARMS)

Most top journals referencing ARMS are indexed by OpenAlex, including several high-DOI outlets such as *Applied Economic Perspectives and Policy* and the *American Journal of Agricultural Economics*. Fewer journals are exclusive to Scopus or Dimensions. This pattern aligns with OpenAlex's dominant coverage of ARMS publications in the previous section.

The Census of Agriculture

Journal coverage for Census-related publications is distributed more evenly across the three sources. Several journals—particularly in environmental and remote sensing fields—are indexed only in Scopus or Dimensions. Shared indexing is common for journals like *Food Policy* and *Agricultural Systems*, helping to explain the high level of overlap between Scopus and Dimensions.

Food Access Research Atlas

This dataset is associated with journals that are broadly indexed across sources. Titles such as the *Journal of Agricultural and Applied Economics* and *Ecological Economics* are covered in all three databases. The strong overlap in journal indexing corresponds with the relatively balanced publication coverage observed in the prior section.

The Food Acquisition and Purchase Survey (FoodAPS)

Many FoodAPS-related journals fall within the nutrition and behavioral sciences domains, and several of these—such as *Appetite* and *Frontiers in Nutrition*—are indexed in OpenAlex. While a subset of journals is also covered by Scopus and Dimensions, OpenAlex appears to index more of the high-volume titles, consistent with its higher share of FoodAPS-related DOIs.

The Household Food Security Survey Module

This dataset draws from a wide range of journals in public health, food policy, and applied economics. Journals such as *Food Security* and *Journal of Nutrition Education and Behavior* appear in all three sources, but some health-focused titles are only indexed in Scopus or Dimensions. These differences likely contribute to the stronger coverage seen in Scopus and Dimensions.

Rural-Urban Continuum Code

Journals citing RUCC span health, epidemiology, and rural development. Many are indexed in Scopus and Dimensions, including *Environmental Research*, *BMC Public Health*, and *Drug and Alcohol Dependence*. OpenAlex has more limited coverage of these titles, consistent with its lower representation of RUCC-related DOIs.

Summary of Journal Coverage by Dataset

Dataset	Dominant Source	Notable Journals Indexed in All Sources	Notable Journals Missing from Some Sources
ARMS	OpenAlex	AJAE, AEPP, Agribusiness	Few missing; OpenAlex covers most top journals
Census of Agriculture	Scopus / Dimensions	Food Policy, Agricultural Systems	Environmental/remote sensing journals missing in OpenAlex
Food Access Research Atlas	Shared	JAAEA, Ecological Economics	Broad overlap; minimal gaps
FoodAPS	OpenAlex	Food Security, Frontiers in Nutrition	Some nutrition journals missing in Scopus/Dimensions
HFSSM	Scopus / Dimensions	JNED, Food Security	Some public health journals missing in OpenAlex
RUCC	Scopus / Dimensions	Environmental Research, Food Policy	Several epidemiology/health titles missing in OpenAlex

3.3 Publication Topics

In addition to differences in coverage and journal indexing, citation databases vary in how they classify research content. Each system applies a distinct taxonomy—often algorithmically generated—to assign topics to publications. These systems function like thematic filters, shaping how research is organized, discovered, and interpreted.

To understand how topic classification differs across sources, this section compares the most frequent topics assigned to the same set of publications by Scopus, OpenAlex, and Dimensions.

Why Focus on Overlapping Publications?

To ensure comparability, the analysis is restricted to DOIs that appear in all three databases. This approach isolates differences in classification by holding the underlying publication set constant. Any observed variation reflects how each database labels and groups the same publications.

Word Cloud Construction

For each dataset, the word clouds are based on frequency tables constructed from topic metadata assigned by each source. Specifically:

- The analysis filters to DOIs indexed by all three sources
- For each source, the corresponding topic classification schema is used to generate a count of how many DOIs are linked to each topic
- The word clouds visualize the top 100 most frequent topics assigned by each source to those shared DOIs

Source-specific classification methods include:

- Scopus: Author keywords and ASJC codes
- OpenAlex: Topic field from OpenAlex's hierarchical ontology
- Dimensions: Concepts assigned using machine learning (per Dimensions API codebook)

A separate frequency table was generated for each source and dataset combination. These topic counts form the basis of the word clouds shown below.

Agricultural Resource Management Survey (ARMS)

For the 35 publications referencing ARMS and indexed in all three sources, topic classification reveals some consistent patterns alongside differences in emphasis across systems.

Dimensions assigns a range of applied and operational topics, including Soil Health, Marketing Channels, Dairy Farms, and Crop Insurance. Several topics focus on production decisions and management practices, such as Farm Size, Data Privacy, and N Leaching. The prominence of

terms related to input use, conservation, and cost structures reflects a policy- and practice-oriented framing.

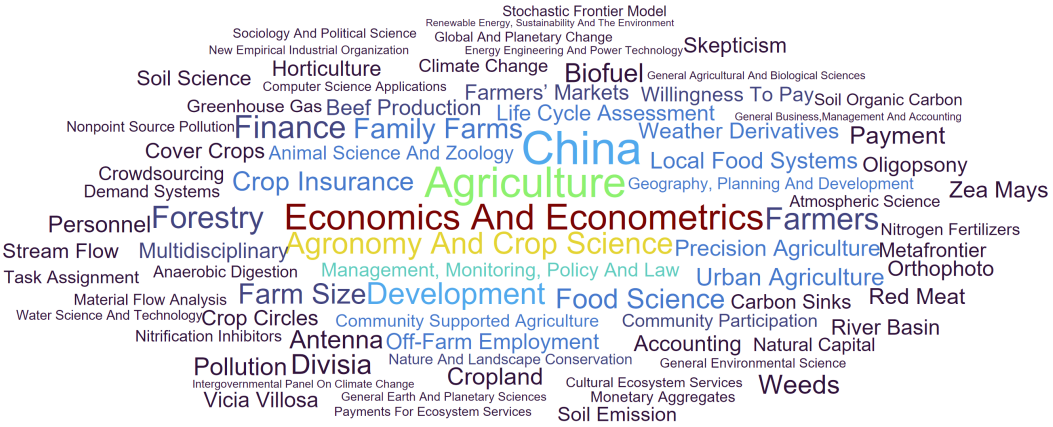
OpenAlex yields a more consolidated view, emphasizing broad thematic categories like Agricultural Innovations and Practices, Organic Food and Agriculture, and Economics of Agriculture and Food Markets. Fewer distinct terms are present, and the focus is more conceptual, with attention to sustainability, risk, and valuation.

Scopus includes disciplinary terms in Economics, Agronomy, and Environmental Science, alongside practical topics such as Farm Size, Cover Crops, and Weather Derivatives. The classification reflects both domain-specific and cross-cutting areas, consistent with the broader journal base in Scopus.

Taken together, these results suggest that while all three sources capture core agricultural themes, Dimensions emphasizes policy tools and production systems, OpenAlex foregrounds conceptual linkages to sustainability and innovation, and Scopus presents a broader disciplinary framing. These differences have implications for how research using ARMS data might be surfaced in topic-based search or evaluation.

3.3.0.1 Scopus

ARMS
Most Frequent Research Topics from Scopus



Only includes DOIs that are indexed in Scopus, OpenAlex, and Dimensions. N = 30 shared DOIs.

3.3.0.2 OpenAlex

ARMS
Most Frequent Research Topics from OpenAlex

Only includes DOIs that are indexed in Scopus, OpenAlex, and Dimensions. N = 30 shared DOIs.

3.3.0.3 Dimensions

ARMS
Most Frequent Research Topics from Dimensions

Only includes DOIs that are indexed in Scopus, OpenAlex, and Dimensions. N = 30 shared DOIs.

ℹ Additional Word Cloud Variants

The Census of Agriculture

The 247 publications referencing the Census of Agriculture and indexed in all three databases provide a rich basis for comparing classification systems. Each source assigns distinct topics to these shared DOIs, revealing differences in how agricultural, environmental, and economic research is categorized.

Dimensions emphasizes applied agricultural practice and land management topics, such as Cover Crops, Food Systems, Land Use, and No-Till. Many terms are operational, with a focus on production practices, conservation techniques, and farm-level outcomes (e.g., Manure, Calf Care, Biosecurity Practices).

OpenAlex presents a broader thematic range, incorporating environmental, behavioral, and economic dimensions. Terms like Urban Agriculture and Sustainability, Economic and Environmental Valuation, Soil Carbon and Nitrogen Dynamics, and Food Waste Reduction suggest a wider scope of inquiry, including sustainability science, agroecology, and interdisciplinary research.

Scopus emphasizes core disciplinary areas such as Ecology, Food Science, Economics and Econometrics, and Agronomy and Crop Science. The word cloud includes scientific and policy-oriented terms like Soil Emission, Pollution, and Land Titling, along with geographic or global framing (e.g., China, Urban Agriculture).

These differences demonstrate how the same set of publications can be interpreted through varied conceptual lenses. Dimensions leans toward practice- and system-level terms, OpenAlex emphasizes sustainability and complexity, while Scopus highlights scientific domains and applied economic constructs.

3.3.0.1 Scopus

Census of Agriculture
Most Frequent Research Topics from Scopus

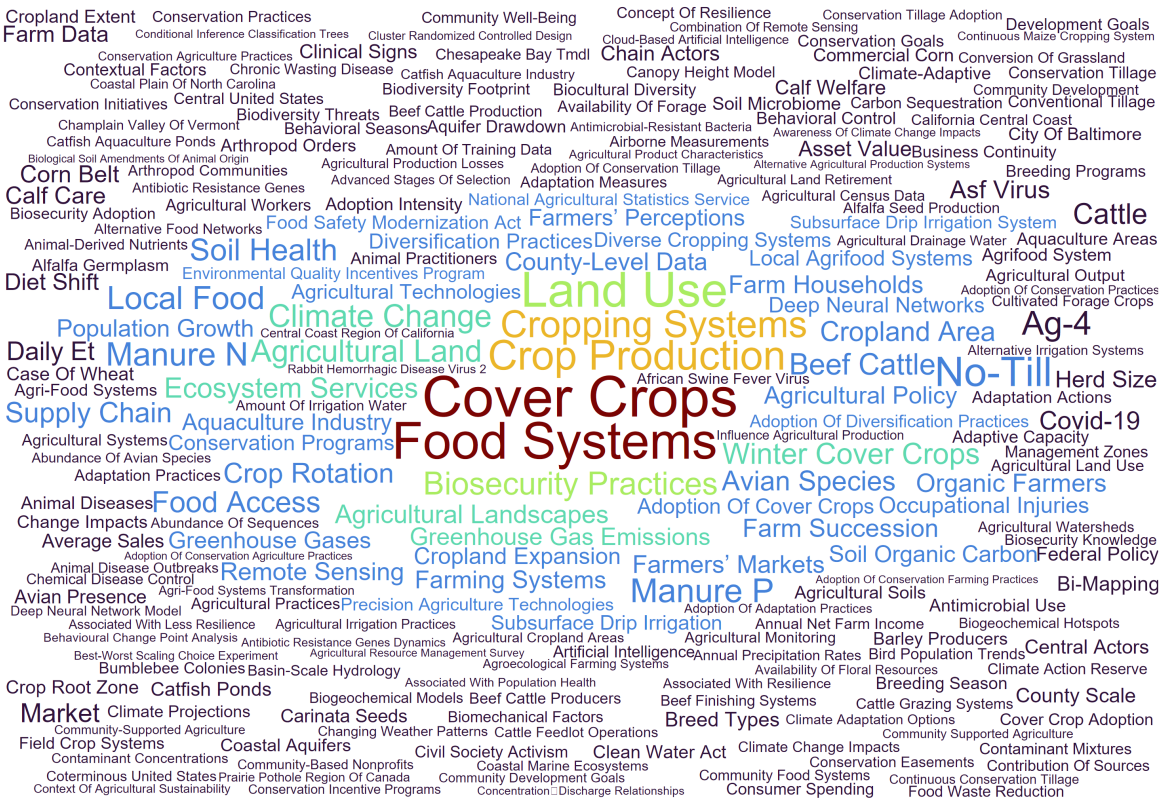
Only includes DOIs that are indexed in Scopus, OpenAlex, and Dimensions. N = 210 shared DOIs.

3.3.0.2 OpenAlex

Census of Agriculture
Most Frequent Research Topics from OpenAlex

Only includes DOIs that are indexed in Scopus, OpenAlex, and Dimensions. N = 210 shared DOIs.

3.3.0.3 Dimensions



Only includes DOIs that are indexed in Scopus, OpenAlex, and Dimensions. N = 210 shared DOIs.

ℹ Additional Word Cloud Variants

Food Access Research Atlas

For the 84 shared DOIs referencing the Food Access Research Atlas, each citation database surfaces a different emphasis in topic classification, reflecting both overlapping and diverging conceptual frames.

Dimensions centers on topics related to Food Deserts, Food Access, and Healthy Food Availability, with strong links to public assistance programs (SNAP, Supplemental Nutrition Assistance Program, Food Insecurity). Additional terms such as Census Tracts, Community Gardening, and Diet Cost point to an applied and community-level framing of food access challenges.

OpenAlex takes a broader thematic approach, highlighting population health topics such as Food Security and Health in Diverse Populations and Obesity, Physical Activity, Diet. There is a noticeable emphasis on intersectional and structural themes—Urban Transport and Accessibility, Health Disparities, Homelessness and Social Issues—alongside sustainability-oriented terms like Urban Agriculture and Sustainability.

Scopus shows a wide spread of disciplinary and technical topics, including Obesity, Grocery Stores, Farmers' Markets, and Public Health. Additional terms such as Anthropology, Surgery, and Roboehtics reflect the influence of journals from medical and applied sciences, sometimes extending beyond food policy and nutrition per se.

Together, these perspectives suggest that Dimensions tends to classify FARA-linked publications in the context of policy and access programs, OpenAlex emphasizes broader social determinants and urban health intersections, while Scopus reflects a more disciplinary and biomedical scope. This variation may shape how researchers and policymakers interpret or retrieve work related to food environments and neighborhood-level food access.

3.3.0.1 Scopus

Food Access Research Atlas
Most Frequent Research Topics from Scopus

Only includes DOIs that are indexed in Scopus, OpenAlex, and Dimensions. N = 65 shared DOIs.

3.3.0.2 OpenAlex

Food Access Research Atlas
Most Frequent Research Topics from OpenAlex

Only includes DOIs that are indexed in Scopus, OpenAlex, and Dimensions. N = 65 shared DOIs.

3.3.0.3 Dimensions

Only includes DOIs that are indexed in Scopus, OpenAlex, and Dimensions. N = 65 shared DOIs.

> Additional Word Cloud Variants

The Food Acquisition and Purchase Survey (FoodAPS)

Among the 45 DOIs referencing FoodAPS that appear in all three sources, each citation database reflects different emphases in how research topics are classified.

Dimensions assigns topics closely aligned with food assistance and economic access, including Diet Cost, Food Environment, Thrifty Food Plan, and Supplemental Nutrition Assistance Program. Many topics reference federal nutrition programs, purchasing behavior, and measures of food affordability, which align with common policy applications of the dataset.

OpenAlex emphasizes broader public health themes, such as Food Security and Health in Diverse Populations and Obesity, Physical Activity, Diet. Related topics highlight structural

and behavioral factors, including Homelessness and Social Issues, Urban Agriculture and Sustainability, and Consumer Attitudes and Food Labeling.

Scopus includes a mix of applied and disciplinary topics, with frequent terms like Obesity, Grocery Stores, Farmers' Markets, and Nutrition and Dietetics. Additional topics such as Brand Placement, Food Labeling, and Program Participation suggest greater representation of behavioral nutrition, labeling policy, and intervention studies.

Overall, Dimensions tends to classify FoodAPS-related work in terms of food policy and program evaluation, OpenAlex highlights public health and social determinants, and Scopus reflects more disciplinary and intervention-focused research.

3.3.0.1 Scopus

Food Acquisition and Purchase Survey (FoodAPS)
Most Frequent Research Topics from Scopus

Only includes DOIs that are indexed in Scopus, OpenAlex, and Dimensions. N = 38 shared DOIs.

3.3.0.2 OpenAlex

Food Acquisition and Purchase Survey (FoodAPS)
Most Frequent Research Topics from OpenAlex

Only includes DOIs that are indexed in Scopus, OpenAlex, and Dimensions. N = 38 shared DOIs.

3.3.0.3 Dimensions

Food Acquisition and Purchase Survey (FoodAPS)
Most Frequent Research Topics from Dimensions

Only includes DOIs that are indexed in Scopus, OpenAlex, and Dimensions. N = 38 shared DOIs.

ℹ Additional Word Cloud Variants

The Household Food Security Survey Module

Among the 98 DOIs indexed in Scopus, OpenAlex, and Dimensions, publications referencing the Household Food Security Survey Module (HFSSM) reflect a wide-ranging set of topics, though the emphasis varies by source.

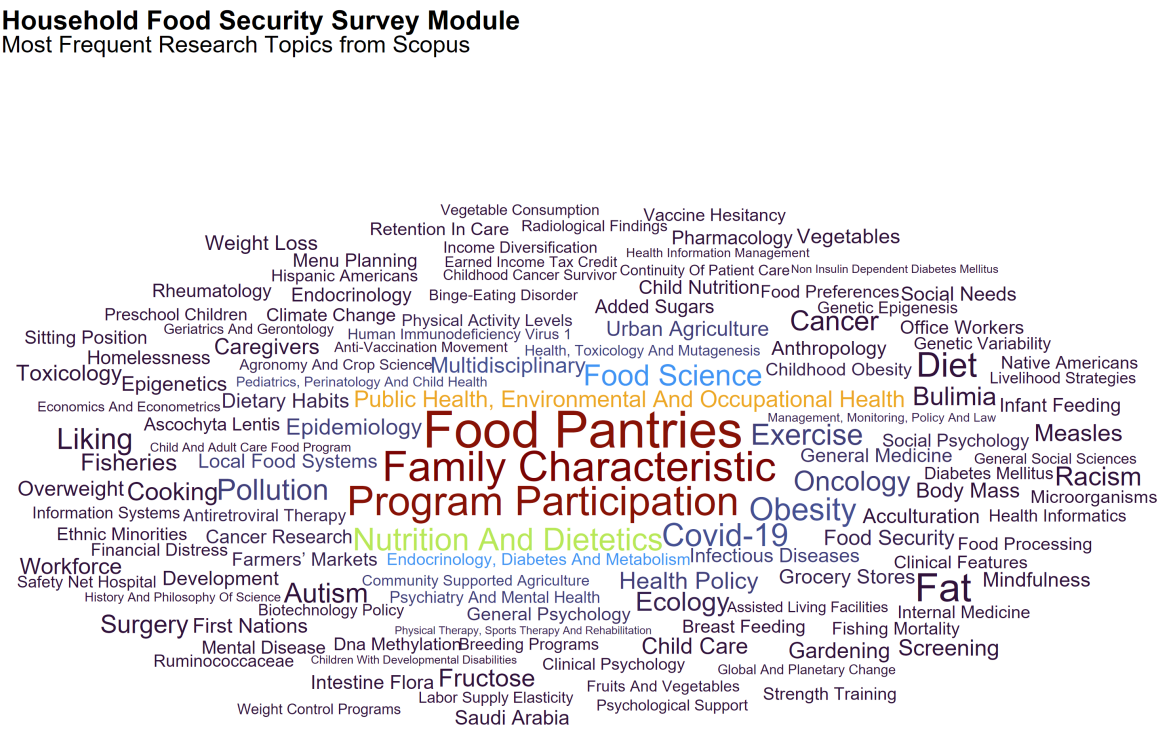
Dimensions associates HFSSM-related publications most frequently with topics such as food insecurity, supplemental nutrition assistance, diet quality, older adults, and mental health. Many of these topics connect to program evaluation, food access, and health outcomes in low-income or vulnerable populations.

Scopus similarly highlights food pantries, program participation, and family characteristics, but also includes a broader range of biomedical and psychological research, with terms like epigenetics, autism, clinical features, and mental disease. This broader disciplinary coverage reflects the indexing structure of Scopus, which spans both health and social science fields.

In OpenAlex, topic classifications are more concentrated. The most common labels include food security and health in diverse populations, obesity, physical activity, and diet, and homelessness and social issues. These reflect a more sociomedical framing and suggest a focus on population-level health disparities and structural determinants.

Taken together, these differences in topical classification illustrate how each database frames HFSSM-related research through its own taxonomic lens. While there is substantial thematic overlap, the terminology and granularity used to describe research topics can shape how a dataset's contributions appear across bibliographic sources.

3.3.0.1 Scopus



Only includes DOIs that are indexed in Scopus, OpenAlex, and Dimensions. N = 82 shared DOIs.

3.3.0.2 OpenAlex

Only includes DOIs that are indexed in Scopus, OpenAlex, and Dimensions. N = 82 shared DOIs.

3.3.0.3 Dimensions

Only includes DOIs that are indexed in Scopus, OpenAlex, and Dimensions. N = 82 shared DOIs.

> Additional Word Cloud Variants

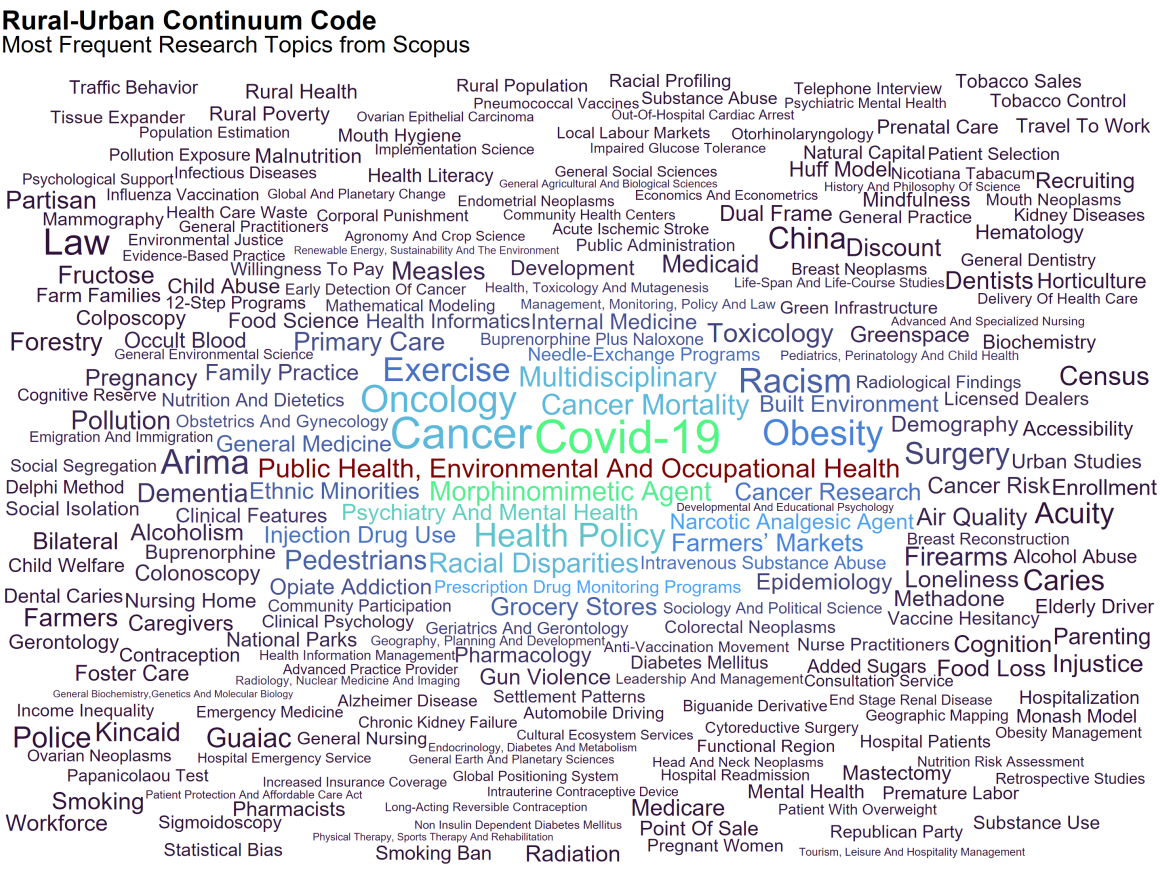
Rural-Urban Continuum Code

There are 164 DOIs referencing RUCC that appear in all three sources. Topic classifications for these publications center on public health, healthcare access, and rural disparities, but the emphasis varies across systems.

Dimensions highlights county-level metrics and rural health infrastructure, with terms like rural counties, older adults, and cancer survivors being especially prominent. OpenAlex places greater weight on population health and health disparities, frequently tagging topics such as opioid use disorder treatment, global cancer incidence, and primary care and health outcomes. In Scopus, classifications reflect a broad span of public health themes—including obesity,

health policy, Medicaid, and exercise—as well as clinical topics like cancer and surgery. These differences reflect each source's unique tagging schema, reinforcing the value of triangulating across systems when analyzing policy-relevant publication domains.

3.3.0.1 Scopus



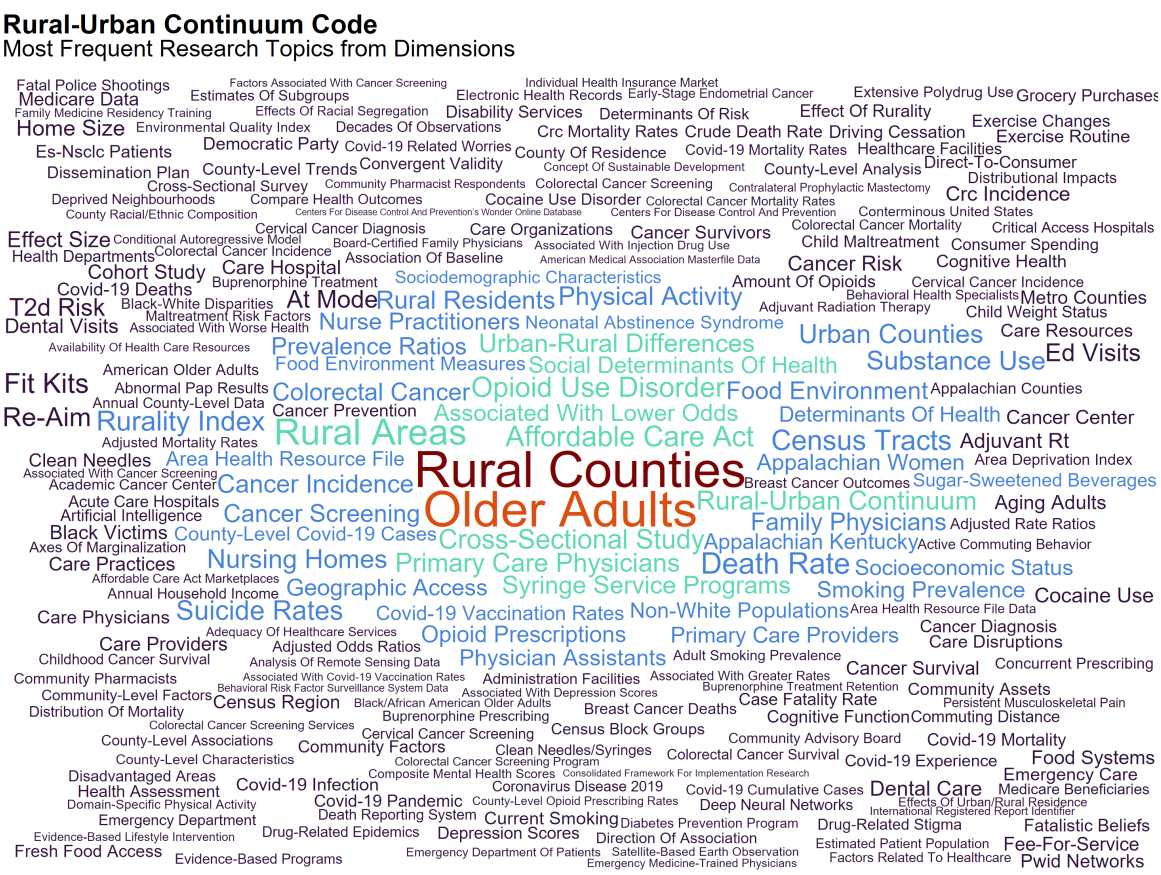
Only includes DOIs that are indexed in Scopus, OpenAlex, and Dimensions. N = 130 shared DOIs.

3.3.0.2 OpenAlex

Rural-Urban Continuum Code
Most Frequent Research Topics from OpenAlex

Only includes DOIs that are indexed in Scopus, OpenAlex, and Dimensions. N = 130 shared DOIs.

3.3.0.3 Dimensions



Only includes DOIs that are indexed in Scopus, OpenAlex, and Dimensions. N = 130 shared DOIs.

ℹ Additional Word Cloud Variants

3.4 Author Comparison

To identify and compare authors across Scopus, OpenAlex, and Dimensions, a multi-step disambiguation process was implemented. Because not all authors have persistent identifiers (e.g., ORCIDs), and because name formatting, use of initials, and institutional affiliations vary across and within sources, a harmonization pipeline was developed. This process follows the structure of the PatentsView disambiguation methodology and includes the following steps:

1. **Name Normalization and Source-Specific Cleaning:** Author names were extracted from each source and cleaned using a consistent normalization function. This involved

transliterating special characters, removing punctuation, standardizing case, and collapsing whitespace. In each database, author records were linked to publication DOIs and enriched with affiliation information where available.

2. **ORCID-Based Canonical Resolution:** When an author's ORCID was present—either directly in OpenAlex or indirectly via Dimensions—it was used as the canonical identifier. ORCID lookups were performed for all DOIs across sources, and a lookup table was constructed to resolve shared authors using both ORCID and cleaned name/DOI matches.

3. **Blocking Using Canopy Construction:** For authors without ORCID identifiers, blocking keys were constructed by combining the first initial and last name to form "canopy" groups. This reduced the number of pairwise comparisons needed for clustering by limiting them to plausible matches.

4. **String Similarity Clustering Within Canopies:** Within each canopy group, Jaro-Winkler string distances were calculated using the cleaned full names. Hierarchical clustering with average linkage was applied, and clusters were formed using a similarity threshold. Each cluster was then assigned a synthetic canonical ID based on the first observed name.

5. **Merging and Source Propagation:** Author mentions across all three sources were merged into a master long-format table, with canonical IDs assigned based on ORCID or string-based clustering. For each publication, flags were added to indicate whether an author appeared in Scopus, OpenAlex, or Dimensions. These flags were propagated to all mentions of a given author within the same DOI.

6. **Institutional Consolidation:** Author affiliations were collapsed across sources by pivoting to a wide format (institution_1, institution_2, etc.) and summarizing into a primary institution field. This structure supported subsequent author-level aggregation and topic classification.

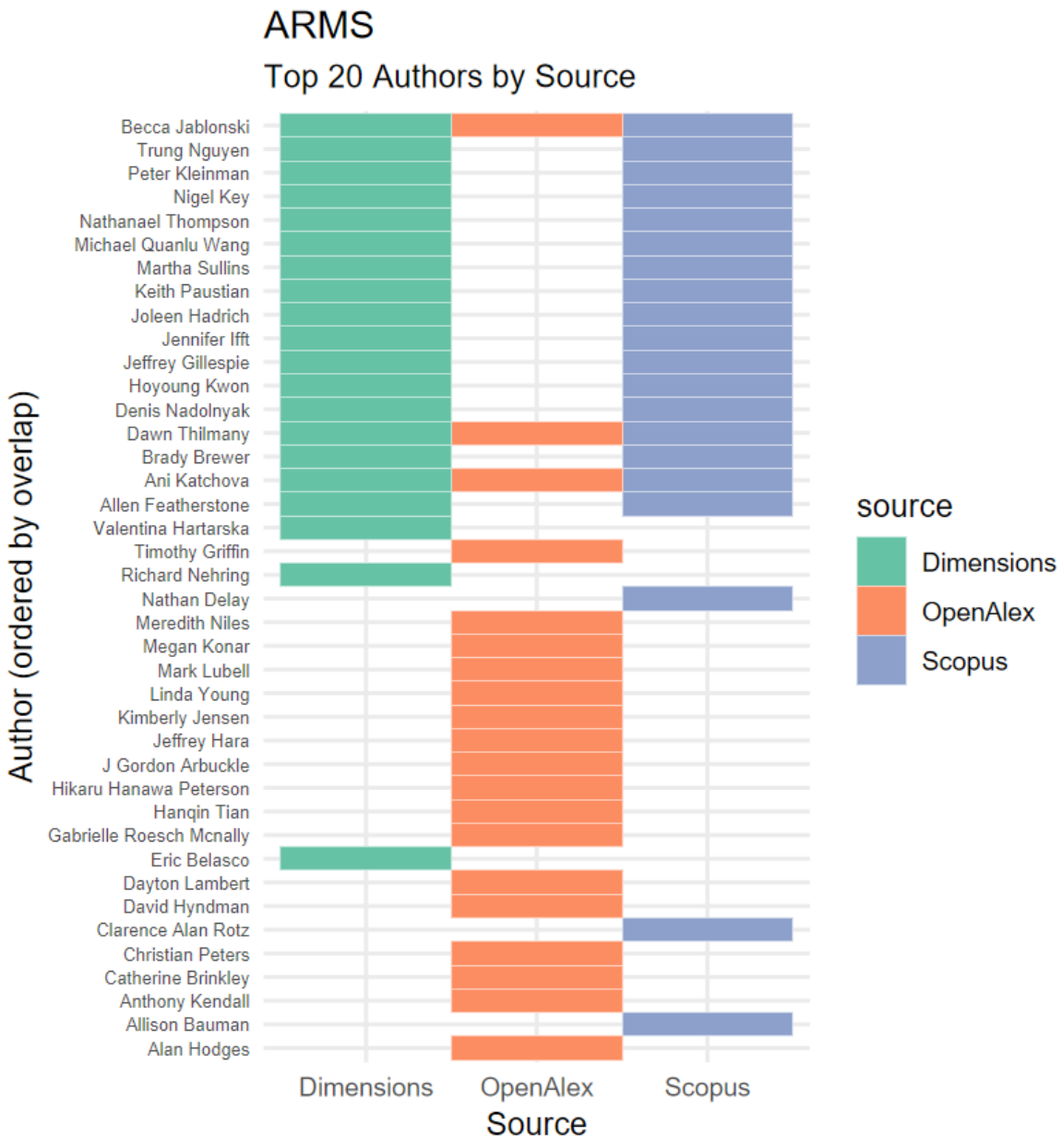
This approach enables the identification of unique authors across bibliometric systems, even in the absence of persistent identifiers. It supports comparisons of author counts, top contributors, and topic-specific participation across Scopus, OpenAlex, and Dimensions.

Main Takeaway

Across all datasets, the authors most visible in one platform are not always discoverable in others. The top contributors to a dataset can vary significantly depending on which citation database is used. These differences stem from inconsistencies in metadata, name disambiguation, and indexing practices. As a result, evaluations or dashboards based on a single source may misrepresent who is using a dataset, leading to undercounting or omission of active researchers. Using multiple sources helps create a more accurate and equitable picture of scholarly engagement.

3.4.1 ARMS

For ARMS, the top 20 authors identified in each platform show limited overlap. While some authors are discoverable across all three sources, others appear only in one, particularly in OpenAlex or Dimensions. This reflects inconsistencies in how author names are indexed and matched across platforms, especially for researchers who publish under multiple name variants or without ORCID identifiers.



This figure shows the top 20 authors by publication count for each source. Differences in author rankings reflect how platform-specific indexing affects who appears as a leading user of a dataset—that is, researchers who most frequently publish work referencing or using it. According to Scopus, this dataset has been used by 734 distinct authors; OpenAlex identifies 4495 distinct users; and Dimensions includes 778.

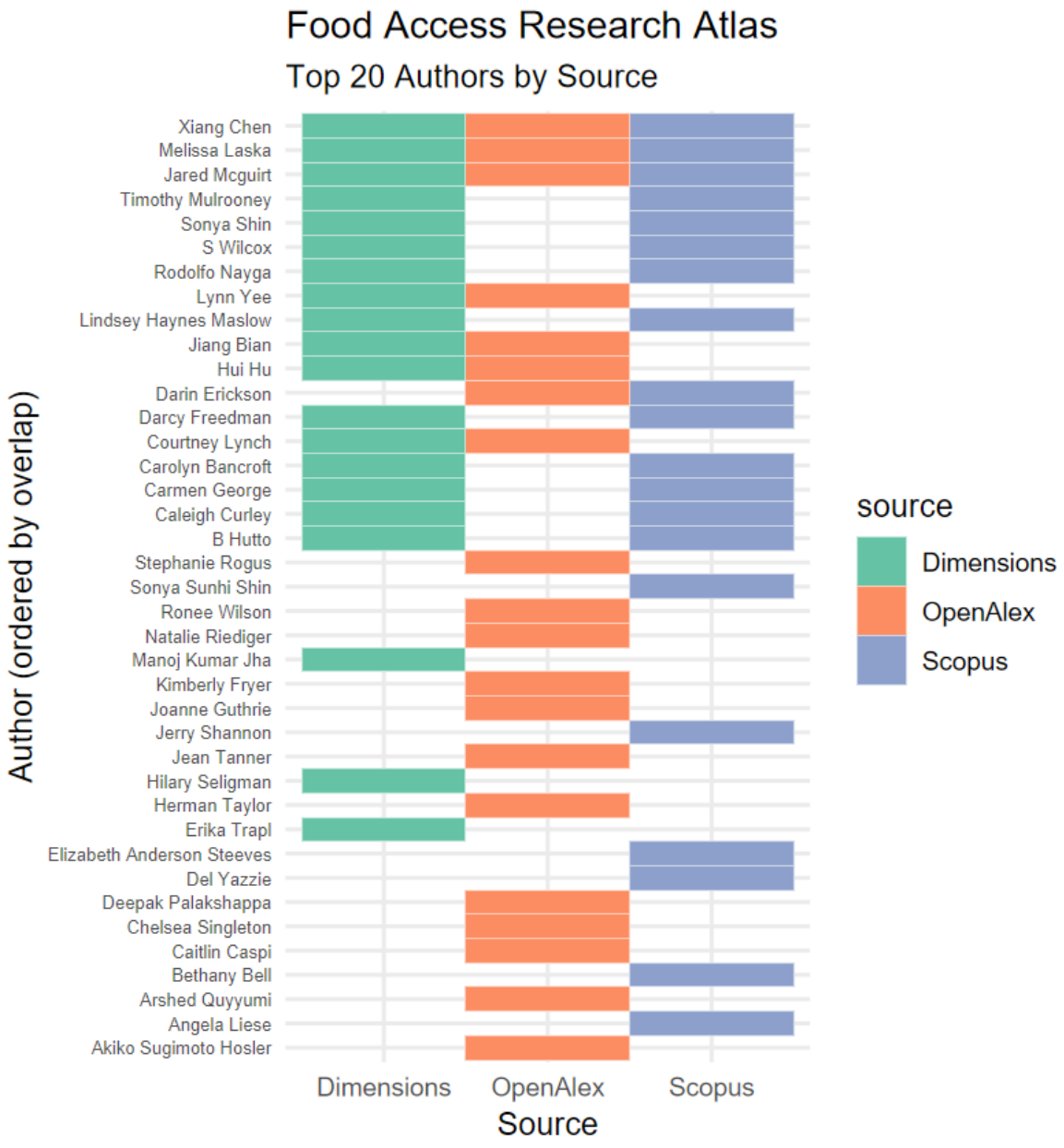
3.4.2 Census of Ag

The Census of Agriculture shows relatively higher agreement across platforms, with many top authors appearing in multiple sources. However, there are still noticeable differences, with some authors ranked highly in one platform but not appearing at all in others. These discrepancies are likely tied to variations in how author metadata and institutional affiliations are recorded.

This figure shows the top 20 authors by publication count for each source. Differences in author rankings reflect how platform-specific indexing affects who appears as a leading user of a dataset—that is, researchers who most frequently publish work referencing or using it. According to Scopus, this dataset has been used by 13351 distinct authors; OpenAlex identifies 4735 distinct users; and Dimensions includes 11332.

3.4.3 FARA

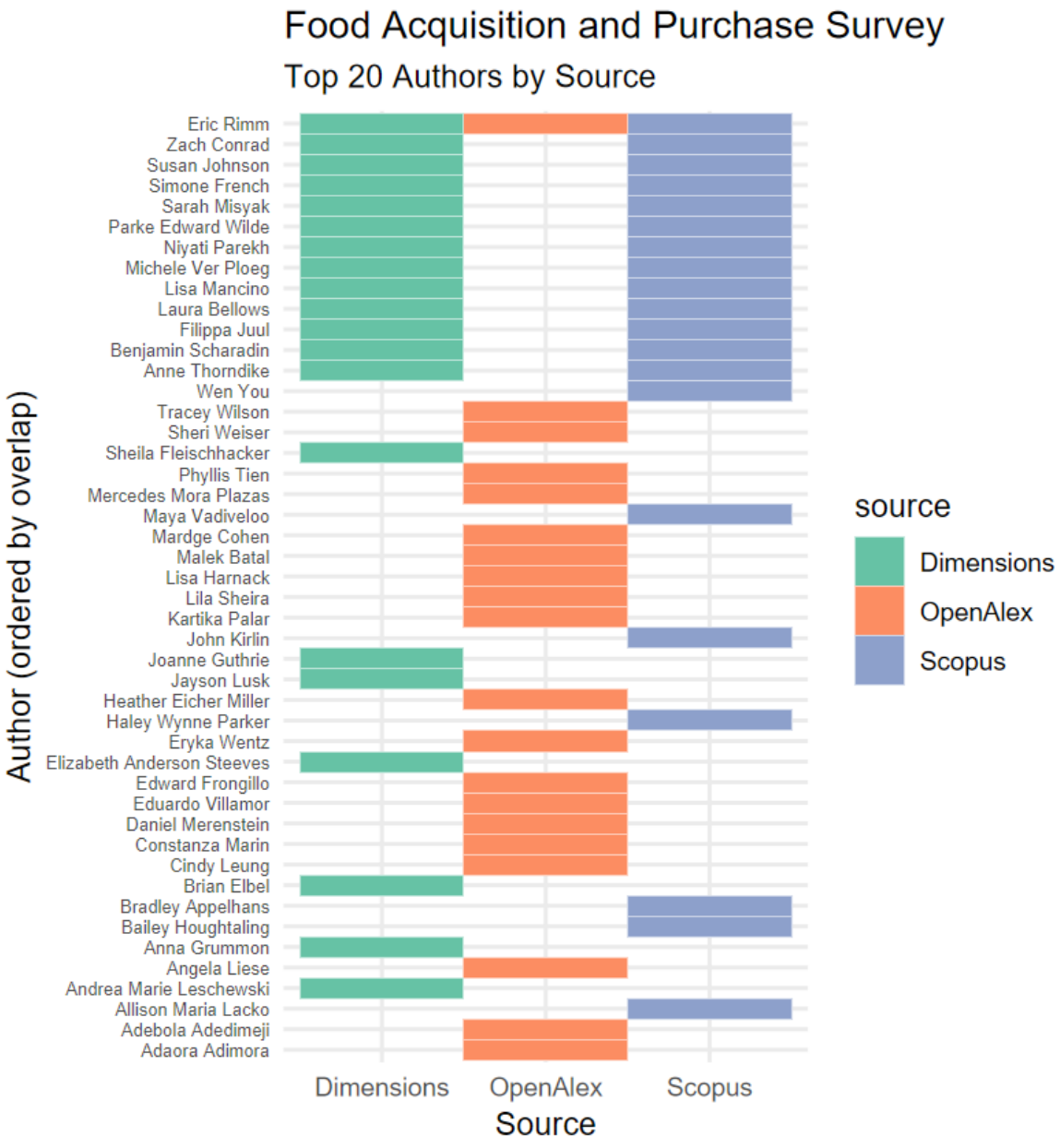
FARA displays substantial divergence in author coverage. A number of top authors are visible in only one platform, and overlap across all three is relatively limited. This dataset seems particularly affected by platform-specific indexing practices—likely because much of the associated research is interdisciplinary and published across a range of journal types.



This figure shows the top 20 authors by publication count for each source. Differences in author rankings reflect how platform-specific indexing affects who appears as a leading user of a dataset—that is, researchers who most frequently publish work referencing or using it. According to Scopus, this dataset has been used by 1610 distinct authors; OpenAlex identifies 1154 distinct users; and Dimensions includes 2013.

3.4.4 FoodAPS

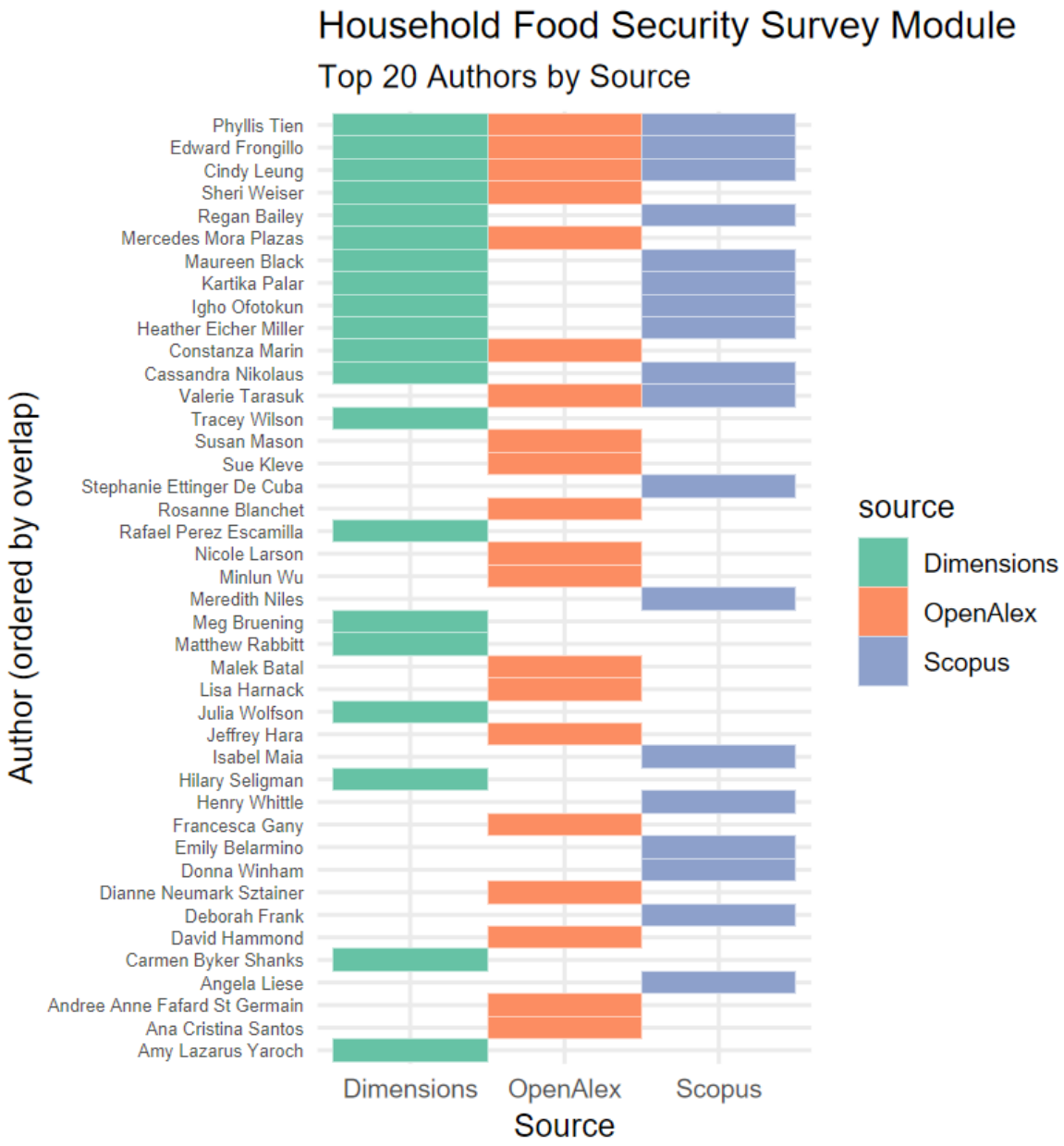
FoodAPS has uneven author coverage across platforms. While some authors are picked up consistently, others are captured by only one source. OpenAlex includes several authors who are not visible in Scopus or Dimensions, suggesting that coverage differences may be especially pronounced for newer researchers or those publishing in open-access venues.



This figure shows the top 20 authors by publication count for each source. Differences in author rankings reflect how platform-specific indexing affects who appears as a leading user of a dataset—that is, researchers who most frequently publish work referencing or using it. According to Scopus, this dataset has been used by 1126 distinct authors; OpenAlex identifies 1849 distinct users; and Dimensions includes 1029.

3.4.5 HFSSM

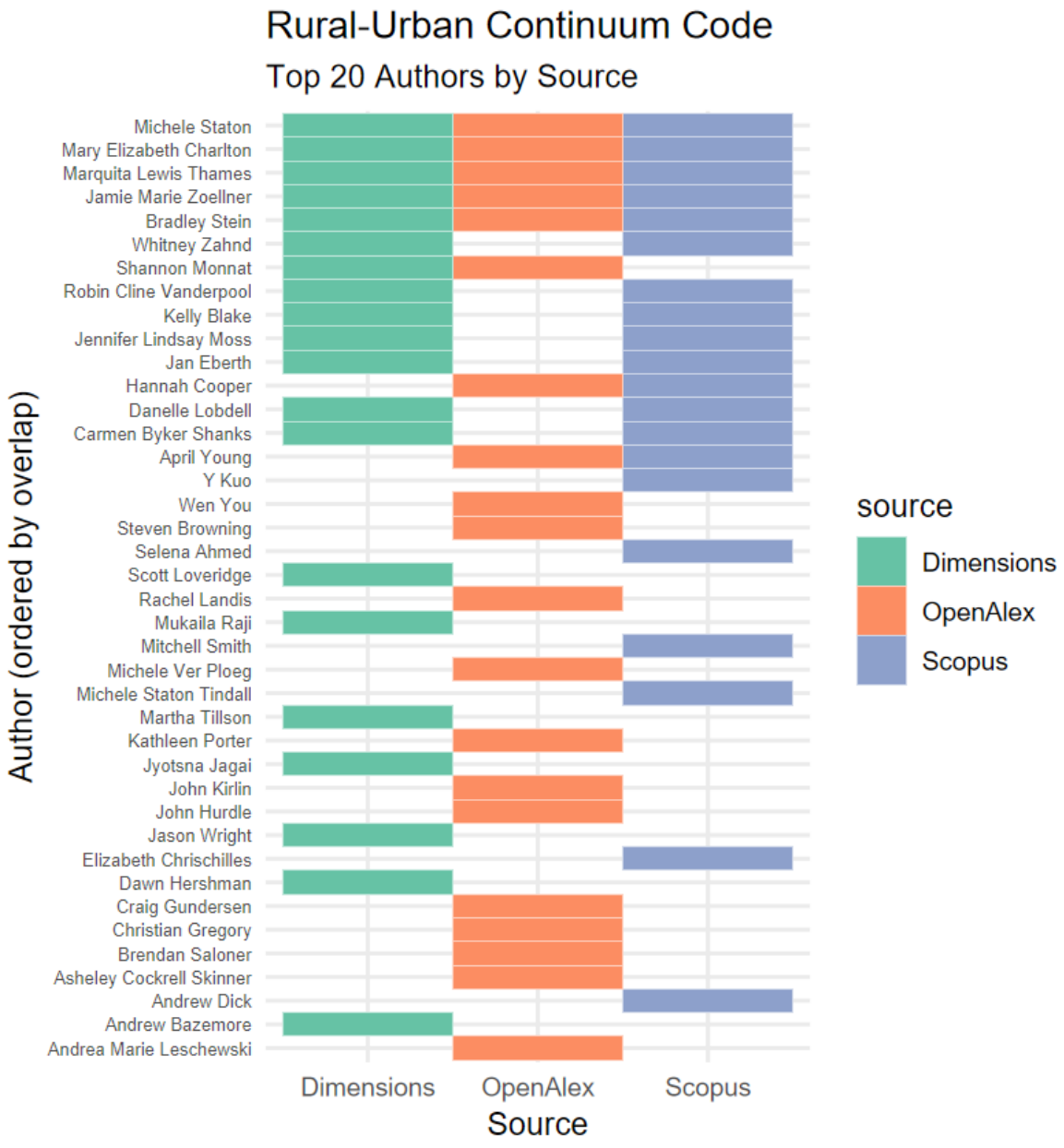
The HFSSM dataset exhibits moderate agreement in author coverage. Most top authors are represented in at least two sources, but each platform still identifies several authors not seen in the others. This suggests that while the dataset has relatively broad exposure, gaps remain that could affect who is counted or highlighted in bibliometric analyses.



This figure shows the top 20 authors by publication count for each source. Differences in author rankings reflect how platform-specific indexing affects who appears as a leading user of a dataset—that is, researchers who most frequently publish work referencing or using it. According to Scopus, this dataset has been used by 3549 distinct authors; OpenAlex identifies 1660 distinct users; and Dimensions includes 3192.

3.4.6 RUCC

RUCC shows the widest variation in author rankings. Many authors appear in only one of the three sources, and very few are consistently represented across all. This fragmentation likely reflects the broad disciplinary scope of RUCC-related research, which spans public health, demography, and social science—fields that are not uniformly indexed across platforms.

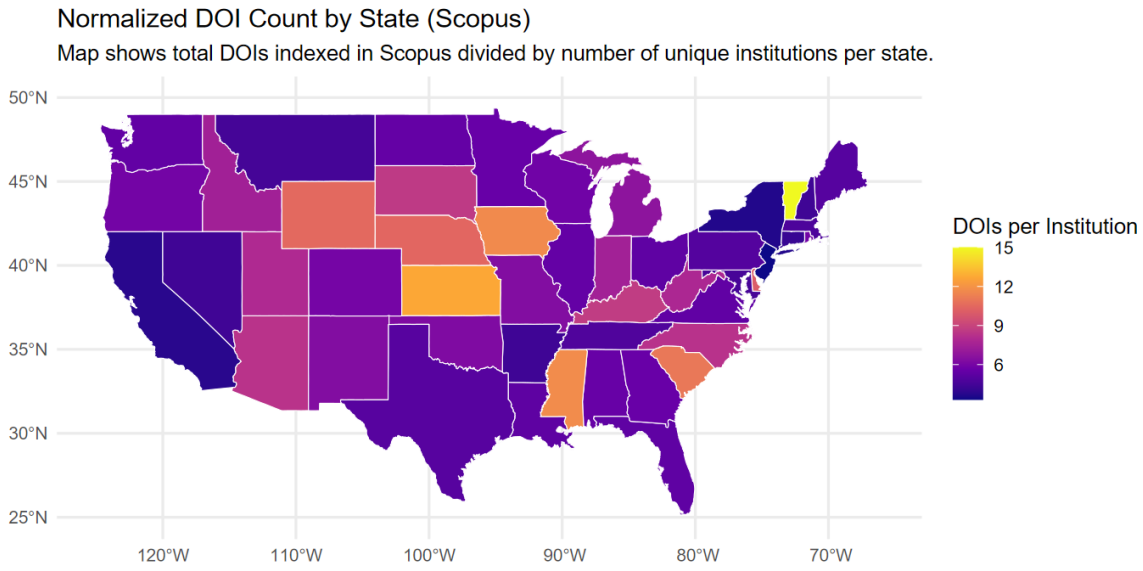


This figure shows the top 20 authors by publication count for each source. Differences in author rankings reflect how platform-specific indexing affects who appears as a leading user of a dataset—that is, researchers who most frequently publish work referencing or using it. According to Scopus, this dataset has been used by 6921 distinct authors; OpenAlex identifies 1881 distinct users; and Dimensions includes 7624.

3.5 Institutional Comparison

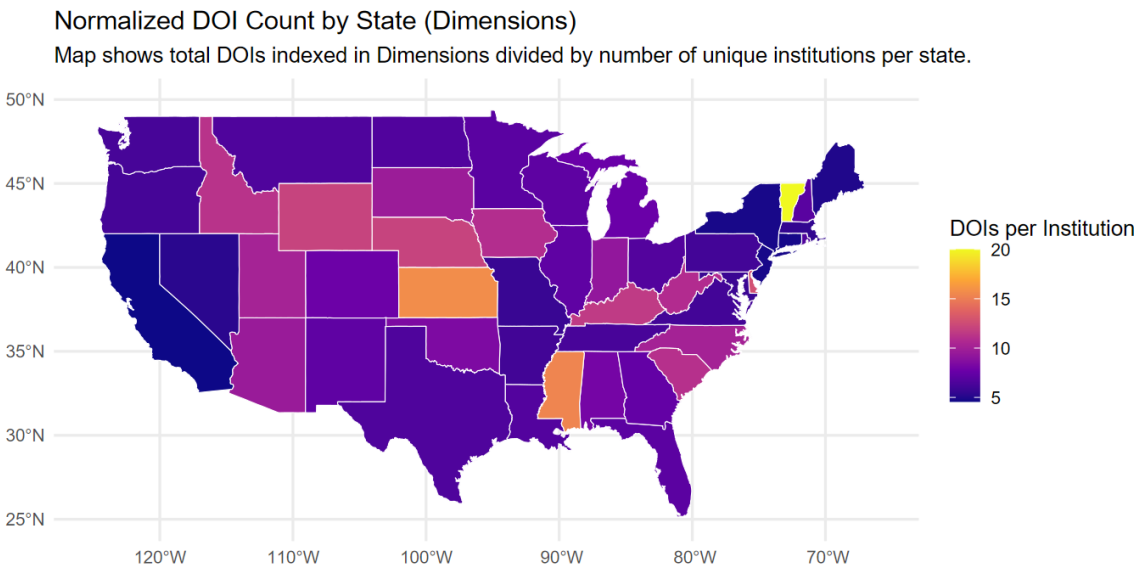
In addition to examining dataset mention coverage, the report also evaluates differences in institutional representation across Scopus, OpenAlex, and Dimensions. Each of the featured citation databases represent some portion of the global research landscape, yet their inclusion criteria and institutional coverage may vary. The purpose of this analysis is to assess which institutions are represented in each source.

3.5.1 Scopus



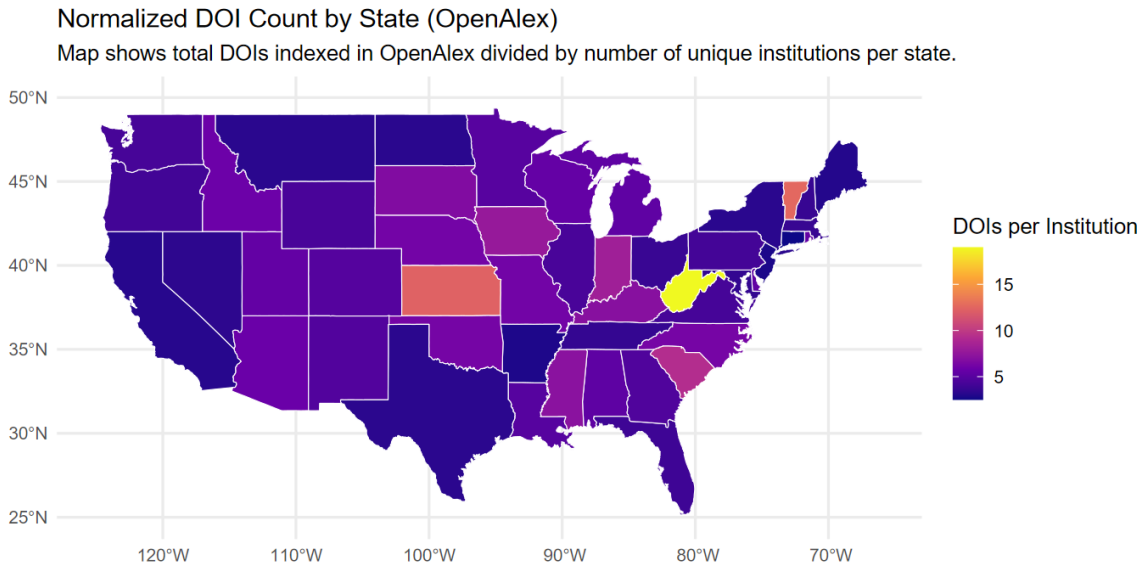
State DOI totals reflect Scopus-indexed publications and are normalized by the number of institutions in each state with at least one publication indexed in this source. How to read: State (total DOIs; total institutions). California (474; 126), District Of Columbia (447; 75), New York (343; 95), North Carolina (339; 41), Maryland (327; 72), Texas (317; 63), Massachusetts (287; 61), Pennsylvania (285; 58), Illinois (272; 50), Virginia (257; 48), Georgia (223; 40), Michigan (221; 33), Florida (211; 40), Ohio (211; 40), Colorado (202; 34), Iowa (186; 16), Wisconsin (186; 32), Minnesota (161; 29), Kentucky (156; 18), Washington (156; 29), Indiana (141; 19), Missouri (139; 22), Tennessee (139; 29), Arizona (133; 16), Kansas (127; 10), Oregon (112; 19), Alabama (105; 19), Nebraska (104; 10), Connecticut (99; 24), South Carolina (99; 9), Utah (95; 12), Arkansas (86; 20), Mississippi (82; 7), Oklahoma (75; 12), Louisiana (68; 13), New Jersey (66; 20), South Dakota (60; 7), Vermont (60; 4), Idaho (59; 8), Montana (58; 13), New Mexico (57; 9), North Dakota (49; 9), Rhode Island (47; 8), Maine (45; 9), New Hampshire (43; 10), Delaware (40; 4), West Virginia (39; 5), Hawaii (23; 6), Wyoming (21; 2), Nevada (13; 3), Alaska (4; 4)

3.5.2 Dimensions



State DOI totals reflect Dimensions-indexed publications and are normalized by the number of institutions in each state with at least one publication indexed in this source. How to read: State (total DOIs; total institutions). District Of Columbia (715; 102), California (648; 142), New York (508; 107), Maryland (499; 87), North Carolina (484; 48), Texas (468; 72), Massachusetts (423; 78), Illinois (415; 58), Pennsylvania (399; 66), Virginia (381; 63), Florida (358; 51), Georgia (338; 46), Michigan (313; 41), Colorado (294; 38), Ohio (287; 43), Wisconsin (271; 38), Indiana (250; 27), Iowa (250; 23), Washington (247; 40), Minnesota (244; 35), Tennessee (237; 38), Arizona (201; 21), Kentucky (196; 17), Missouri (187; 33), Nebraska (180; 15), Kansas (173; 11), Oregon (165; 27), Alabama (161; 20), South Carolina (154; 14), Connecticut (146; 31), Mississippi (138; 9), Oklahoma (134; 16), Utah (132; 13), Arkansas (120; 20), New Jersey (109; 23), Louisiana (107; 16), Vermont (100; 5), South Dakota (87; 9), New Mexico (79; 11), Idaho (78; 7), Montana (78; 12), New Hampshire (69; 10), North Dakota (66; 10), Maine (59; 12), Rhode Island (59; 9), West Virginia (54; 5), Delaware (51; 4), Hawaii (39; 6), Wyoming (36; 3), Nevada (26; 5), Alaska (13; 4)

3.5.3 OpenAlex



State DOI totals reflect OpenAlex-indexed publications and are normalized by the number of institutions in each state with at least one publication indexed in this source. How to read: State (total DOIs; total institutions). California (392; 128), District Of Columbia (385; 76), New York (298; 94), Maryland (289; 83), North Carolina (238; 38), Massachusetts (216; 59), Pennsylvania (205; 50), Illinois (198; 46), Texas (186; 58), Georgia (185; 40), Virginia (184; 44), Michigan (174; 33), Wisconsin (160; 29), Colorado (155; 32), Florida (155; 40), Washington (140; 33), Ohio (137; 37), Minnesota (133; 27), Iowa (131; 17), Indiana (122; 15), Tennessee (119; 34), Missouri (104; 17), Kansas (98; 8), Kentucky (98; 14), Nebraska (92; 15), Arizona (87; 15), Oregon (85; 21), Oklahoma (81; 13), Alabama (79; 15), South Carolina (74; 8), Vermont (63; 5), Connecticut (60; 24), Utah (60; 11), Mississippi (57; 8), New Jersey (53; 19), Arkansas (48; 17), South Dakota (47; 7), Louisiana (44; 9), Idaho (41; 7), Montana (38; 12), New Hampshire (36; 10), New Mexico (33; 7), Rhode Island (33; 6), Maine (27; 9), North Dakota (19; 6), West Virginia (19; 1), Delaware (16; 3), Hawaii (13; 6), Nevada (13; 4), Wyoming (13; 3), Alaska (4; 2)

4 Conclusion

This report compares the coverage of publications and journals referencing the Census of Agriculture across Scopus and OpenAlex, using two approaches for identifying relevant OpenAlex publications: a full-text search and a seed corpus approach.

Using the full-text search in OpenAlex, we found relatively limited overlap with Scopus. Only 9.2% of publications and 9.2% of journals referencing the Census of Agriculture appeared in both databases, with Scopus identifying a substantially larger share of relevant works. These results suggest that relying solely on OpenAlex's full-text search may miss a significant number of dataset mentions.

Applying the seed corpus approach to OpenAlex improved overlap with Scopus and provided a more structured way to capture publications associated with known journals, authors, and

topics. However, the percentage of overlapping publications referencing the Census of Agriculture is lower at 6.42% even though there is a slightly higher percentage of shared journals at 10.73%.

Comparing the overlap between the two OpenAlex methods reveals differences in underlying samples. Only 20.8% of full-text search publications were also found in the seed corpus set, and 28.9% of seed corpus publications matched those found in the full-text search. Journal-level overlap was somewhat higher, with 137 journals shared between the two methods (representing approximately 50–55% overlap across the two pools).

It is important to note that the full-text search and seed corpus approaches represent two distinct sampling methods within OpenAlex. The full-text search attempts to identify dataset mentions directly from the body of text available for a subset of publications, while the seed corpus approach relies on pre-selected journals, topics, and authors more likely to reference the Census of Agriculture. As a result, the pools of publications identified by each method are not strictly comparable: they are drawn from different underlying subsets of OpenAlex's catalog. This context is important for interpreting differences in coverage and citation intensity across the two approaches.

Tables

Table 6: Top 25 Topics by First Run Count

Topic ID	Topic Name	Full-Text Search Count	Total Count
T11610	Impact of Food Insecurity on Health Outcomes	549	78661
T10010	Global Trends in Obesity and Overweight Research	272	111686
T11066	Comparative Analysis of Organic Agricultural Practices	247	41275
T12253	Urban Agriculture and Community Development	222	27383
T10367	Agricultural Innovation and Livelihood Diversification	186	49818
T11464	Impact of Homelessness on Health and Well-being	175	101019
T12033	European Agricultural Policy and Reform	137	88980
T10841	Discrete Choice Models in Economics and Health Care	126	66757
T10596	Maternal and Child Nutrition in Developing Countries	116	118727
T11898	Impacts of Food Prices on Consumption and Poverty	113	29110
T11259	Sustainable Diets and Environmental Impact	109	45082
T11311	Soil and Water Nutrient Dynamics	84	52847
T10235	Impact of Social Factors on Health Outcomes	81	86076
T10439	Adaptation to Climate Change in Agriculture	77	27311
T11886	Risk Management and Vulnerability in Agriculture	73	44755
T10226	Global Analysis of Ecosystem Services and Land Use	71	84104
T10866	Role of Mediterranean Diet in Health Outcomes	70	76894
T10969	Optimal Operation of Water Resources Systems	70	97570
T10330	Hydrological Modeling and Water Resource Management	69	132216
T11753	Forest Management and Policy	60	75196
T12098	Rural development and sustainability	54	62114
T10111	Remote Sensing in Vegetation Monitoring and Phenology	52	56452
T10556	Global Cancer Incidence and Mortality Patterns	49	64063

Topic ID	Topic Name	Full-Text Search Count	Total Count
T11711	Impacts of COVID-19 on Global Economy and Markets	49	69059
T12724	Integrated Management of Water, Energy, and Food Resources	47	40148

Table 7: Top 25 Journals by First Run Count

Journal ID	Journal Name	Full-Text Search Count	Total Count
S2764628096	Journal of Agriculture Food Systems and Community Development	57	825
S115427279	Public Health Nutrition	51	3282
S206696595	Journal of Nutrition Education and Behavior	41	3509
S15239247	International Journal of Environmental Research and Public Health	39	59130
S4210201861	Applied Economic Perspectives and Policy	39	647
S10134376	Sustainability	35	87533
S5832799	Journal of Soil and Water Conservation	34	556
S2739393555	Journal of Agricultural and Applied Economics	34	329
S202381698	PLoS ONE	30	143568
S124372222	Renewable Agriculture and Food Systems	30	426
S200437886	BMC Public Health	28	18120
S91754907	American Journal of Agricultural Economics	28	876
S18733340	Journal of the Academy of Nutrition and Dietetics	27	5301
S78512408	Agriculture and Human Values	27	938
S110785341	Nutrients	25	30911
S2764593300	Agricultural and Resource Economics Review	25	247
S4210212157	Frontiers in Sustainable Food Systems	23	3776
S63571384	Food Policy	20	1069
S69340840	The Journal of Rural Health	20	749
S4210234824	EDIS	18	3714
S19383905	Agricultural Finance Review	18	327
S119228529	Journal of Hunger & Environmental Nutrition	17	467
S43295729	Remote Sensing	14	33899
S2738397068	Land	14	9774
S80485027	Land Use Policy	14	4559

Table 8: Top 25 Authors by First Run Count Table

Author ID	Author Name	Full-Text Search Count	Total Count
A5016803484	Heather A. Eicher-Miller	15	140
A5024975191	Edward A. Frongillo	13	351
A5055158106	Becca B.R. Jablonski	12	60
A5047780964	Meredith T. Niles	11	200
A5076121862	Sheri D. Weiser	10	241
A5068812455	Cindy W. Leung	10	170
A5062679478	J. Gordon Arbuckle	10	68
A5015017711	Jeffrey K. O'Hara	10	27
A5081656928	Whitney E. Zahnd	9	147
A5002438645	Phyllis C. Tien	8	244
A5035584432	Angela D. Liese	8	172
A5027684365	Dayton M. Lambert	8	110
A5081012770	Linda J. Young	8	51
A5008463933	Catherine Brinkley	8	34
A5030548116	Michele Ver Ploeg	8	33
A5056021318	Nathan Hendricks	7	320
A5024248662	Adebola Adedimeji	7	137
A5002732604	Julia A. Wolfson	7	137
A5038610136	Christopher N. Boyer	7	115
A5044317355	Daniel Merenstein	7	113
A5006129622	Carmen Byker Shanks	7	103
A5060802257	Tracey E. Wilson	7	102
A5050792105	Jennifer L. Moss	7	90
A5032940306	Lisa Harnack	7	89
A5024127854	Eduardo Villamor	7	84