Methodology for Comparing Citation Database Coverage of Dataset Usage

Findings

2025-06-04

Table of contents

Re	port	Summary	3
	Wha	at Is the Issue?	3
	How	Was the Study Conducted?	4
	Wha	at Did the Study Find?	5
	How	to Use This Report	6
1	Proj	ect Background	7
	1.1	Project Objective	7
	1.2	Specific Aims	8
2	Data	a Collection	9
	2.1	Scopus Approach	0
	2.2	OpenAlex Approach	1
		2.2.1 OpenAlex Full-Text Search Approach	1
		2.2.2 OpenAlex Seed Corpus Approach	3
	2.3	Dimensions	5
3	Resi	ılts 1	5
	3.1	Publication Coverage	6
		ARMS Financial and Crop Production Practices	7
		The Census of Agriculture	8
		Food Access Research Atlas	9
		The Food Acquisition and Purchase Survey (FoodAPS)	0
		The Household Food Security Survey Module	1
		Rural-Urban Continuum Code	2

Con	clusion	25
	3.4.2 Results of Institutional Comparision	25
	3.4.1 IPEDS and MSI Classifications	24
3.4	Author and Institutional Comparison	23
	Rural-Urban Continuum Code	23
	The Household Food Security Survey Module	23
	The Food Acquisition and Purchase Survey (FoodAPS)	23
	Food Access Research Atlas	23
	The Census of Agriculture	23
	ARMS Financial and Crop Production Practices	23
3.3	Publication Topics	23
	Rural-Urban Continuum Code	23
	The Household Food Security Survey Module	23
	The Food Acquisition and Purchase Survey (FoodAPS)	23
	Food Access Research Atlas	23
	The Census of Agriculture	23
	ARMS Financial and Crop Production Practices	23
3.2	Journal Coverage	22

4 Conclusion

Tables

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.

¹A 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 **Add Rafael's text**. 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)
- 6. Household Food Security Survey Module (HHFSS)

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

³The 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.

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

The objective of this project is to compare publication coverage across citation databases to determine how data usage varies depending on the database. These findings inform decisions about data preservation and future data investments,

To inform this decision, we compare the coverage, structure, and metadata quality of three citation databases—Scopus, OpenAlex, and Dimensions—focusing on their ability to support consistent and transparent dataset usage metrics across the research landscape.

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. 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.
- 5. Evaluate author representation. Compare how author names are recorded across platforms, including the completeness of author metadata and potential implications for attribution and visibility.
- 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.

These aims 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 institutional affiliation: Determining how each platform records institutional information

The scope of work includes comparing publication coverage across Scopus, OpenAlex, and Dimensions that mention select USDA datasets. This inclusion provides a comprehensive assessment of citation databases, particularly in evaluating dataset coverage across both proprietary and open-access platforms. For more information on each citation database, refer to this Appendix. The methodology described in this report provides a systematic approach for comparing publication coverage where federal datasets are mentioned across citation databases. 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.

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.

Table I. List of CSD11 Data Hobers	Table	1:	List	of	USDA	Data	Assets
------------------------------------	-------	----	------	----	------	------	--------

Dataset Name	Produced By	Description
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.

To provide a comprehensive reference for dataset tracking, this Appendix includes a detailed list of data assets and their corresponding aliases, collectively referred to as *dyads*. Each dyad represents a dataset-name and alias pair used in citation database searches, allowing for more precise identification of dataset mentions in research publications. These aliases include acronyms, alternate spellings, dataset variations, and associated URLs, ensuring broad coverage across different citation practices. The dyad list serves as the foundation for dataset extraction and disambiguation across Scopus, OpenAlex, and Dimensions.

To identify relevant publications for each of the six datasets, various search strategies were used across the citation databases: a seed search in Scopus, a full-text metadata search in OpenAlex as well as 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 the Census of Agriculture. 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 selected datasets, including the Census of Agriculture, 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 reference the Ag Census. 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 Census of Agriculture. 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. OpenAlex offers public access to metadata and, when available, full-text content for open-access publications through its API. Unlike Scopus, which provides controlled access to licensed content, OpenAlex indexes only publications that are openly available or for which open metadata has been provided by publishers.

For OpenAlex, two approaches were used to identify publications referencing the Census of Agriculture. The first approach relied on a full-text search across OpenAlex publication records. The second approach applied a seed corpus methodology, similar to the strategy used for Scopus, to address limitations observed in the initial full-text search. In both approaches, the analysis was restricted to articles only. This decision reflects the open-access structure of OpenAlex, where multiple versions of the same work, such as preprints and accepted manuscripts, may be publicly available. To improve consistency and avoid duplication, we include only the final published version.

2.2.1 OpenAlex Full-Text Search Approach

The methodology for collecting mentions of USDA datasets in OpenAlex relied on constructing search queries that combined dataset "aliases" and associated "flag terms" within the text of scholarly works. Dataset aliases represented alternative ways researchers refer to a dataset, such as variations on the Census of Agriculture's official name. Flag terms represented the institutions or agencies responsible for maintaining the dataset. The combination of dataset alias and flag terms ensured that retrieved publications made an explicit connection to the correct data source. A mention was recorded only if at least one alias and one flag term appeared in the same publication, thereby increasing the likelihood of capturing genuine dataset references rather than incidental matches to individual words.⁴

To implement these searches efficiently, the OpenAlex API was accessed using the pyalex Python package.⁵

Search queries were constructed based on OpenAlex's public API documentation, using both the "Filter Works" and "Search Works" endpoints. Filtering parameters were applied to restrict results to English-language publications, published after 2017, classified as articles or reviews, and available through open-access sources.

Boolean logic was used to define the text search structure. For example, the query for the Census of Agriculture grouped several dataset aliases, including "Census of Agriculture," "USDA Census of Agriculture," "Agricultural Census," and "USDA Census." These aliases were combined using an OR operator. Separately, flag terms including "USDA," "U.S. Department of Agriculture," "United States Department of Agriculture," "NASS," and "National Agricultural Statistics Service" were also grouped using an OR operator. The final query ensured that both an alias and a flag term appeared by connecting the two groups with an AND operator:

("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")

This structure required that each publication mention both a recognized variant of the Census of Agriculture name and a reference to the institution responsible for producing it.

Publications matching the query were returned in JSON format, based on the OpenAlex "Work object" schema. 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)
- cited_by_count (citation counts)
- type (publication type, e.g., "article")

⁴Initial 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.
⁵Pyalex 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.

- publication_year (year article was publish)
- language (language, English only)
- is_oa (open access)

Although a range of publication types were retrieved—including articles, book chapters, dissertations, preprints, and reviews—approximately 80–85 percent were classified as articles. To standardize the dataset for downstream analysis, results were filtered during the search process to retain only records identified as type = article. This step removed preprints and non-final versions of works, supporting a more standardized analysis of dataset mentions in peer-reviewed literature.

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

2.2.1.1 Limitations of OpenAlex Full-Text Approach

Although the OpenAlex API provides full-text search capabilities, limitations in how publication content is ingested and indexed introduce challenges for identifying dataset mentions accurately.

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 the Census of Agriculture and other USDA datasets that appear either exclusively in references or are fragmented within n-grams may be missed. To address these limitations, an alternative search strategy was developed based on constructing a filtered seed corpus of publications for local full-text analysis.

2.2.2 OpenAlex Seed Corpus Approach

To address limitations in OpenAlex's full-text indexing methods, a seed corpus approach was applied. The objective was to create a filtered set of publications for local text search to better capture dataset mentions.

To construct the seed corpus, publications were filtered based on several criteria:

- Language: English
- Publication Year: 2017-2023
- Publication Type: Articles only
- Open Access Status: Open-access publications only

Filtering was further refined by selecting publications associated with high-relevance topics, journals, and authors. As an example, the tables shown for the Census of Agriculture dataset— Table 3 (top 25 topics), Table 4 (top 25 journals), and Table 5 (top 25 U.S.-affiliated authors) illustrate how this filtering process was applied. Each table presents two key columns to support interpretation. The *First Run Count* refers to the number of publications linked to each entity (whether a topic, journal, or author) based on metadata from OpenAlex's full-text search feature. This count reflects how often USDA datasets were mentioned within the full text of publications associated with a particular entity. The *OpenAlex Total Count* represents the total number of publications linked to that entity in the broader OpenAlex database, without applying any filters related to dataset mentions.

To create a more focused and manageable search corpus, we selected the top 25 entities in each category based on their First Run Count. This approach prioritizes journals, topics, and authors where USDA datasets are most frequently mentioned in the full text, which we interpret as being more representative of actual research activity involving these datasets. It also substantially reduces the workload by limiting the number of publications that need to be retrieved and processed. The results of this search generated a set of json files. The Python script used to flatten OpenAlex Seed Corpus JSON files is provided in this Appendix. **UPDATE LINK**

Choosing this approach has a few important implications. First, it likely increases the relevance of the resulting corpus by concentrating on publications where USDA data are actively cited or discussed, rather than simply associated with a broader research area. Second, it helps avoid the need to download and process an unmanageable number of PDFs—estimated at around 1.7 million if all identified entities were included. However, this method may introduce some selection bias by favoring entities with higher immediate visibility in the first search pass. Some relevant but less frequently mentioned entities might be excluded, meaning that while efficiency improves, full comprehensiveness is slightly sacrificed. Overall, this trade-off supports a practical balance between depth and feasibility in building the final dataset of publication metadata.

For the Census of Agriculture, the resulting seed corpus included approximately 1,774,245 unique publications. An initial download of full texts achieved a success rate of roughly 35%, corresponding to an estimated 625,000 accessible full-text documents. Local full-text searches were conducted on this subset to improve detection of dataset mentions beyond what was possible through OpenAlex's built-in search capabilities.

Although the seed corpus approach allows for a more targeted retrieval, limitations remain. Full-text download success was constrained by incomplete or inaccessible open-access links, and processing the entire corpus was computationally intensive. Future efforts may require distributed processing or refined selection criteria to further improve efficiency.

Results from both methods are compared to assess differences in dataset mention detection across approaches.

2.3 Dimensions

Coming Soon [Rafael]

In the Dimensions dataset, publication records are classified under several document types, including "article", "chapter", "preprint", "proceeding", and "monograph". For the purposes of this comparative report, only records classified as "article" are included. This restriction maintains consistency with the Scopus and OpenAlex approaches and helps minimize duplication across versions of the same work.

Additional filtering criteria, such as publication year, language, and open-access status, will be documented once the data extraction and cleaning processes are complete.

3 Results

To produce a consistent count of unique publications referencing each USDA dataset, we consolidated records from three sources-Scopus, OpenAlex (*separately for* Full Text and Seed Corpus approaches), and Dimensions-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.

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 2.

Dataset Name	Number of Unique Publications
ARMS	1,611
Census of Agriculture	6,206
Food Access Research Atlas	639
Food Acquisition and Purchase Survey	839
Household Food Security Survey Module	1,537
Rural-Urban Continuum Code	2,431

Table 2: Unique Publications with Metadata across Sources

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 and OpenAlex. 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 featured in this study, a treemap visualization is presented to summarize publication coverage across the citation databases. Because two search strategies were applied to OpenAlex, the results distinguish publications identified separately. Each treemap groups publications into mutually exclusive categories based on their presence in one or more of the data sources: Scopus, OpenAlex Full Text, OpenAlex Seed Corpus, Dimensions. The size of each box is proportional to the number of distinct DOIs in that group, providing a visual summary of the relative coverage across sources. For example, a large "Scopus only" segment indicates a high number of publications indexed exclusively in Scopus, while overlapping segments (e.g., "Scopus OA Seed") reflect shared coverage between platforms.

⁶In 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.

ARMS Financial and Crop Production Practices

Publication Coverage by Source for ARMS Total Distinct DOIs: 1611



The Census of Agriculture

Publication Coverage by Source for Census of Agriculture Total Distinct DOIs: 6206



Food Access Research Atlas

Publication Coverage by Source for Food Access Research Atlas Total Distinct DOIs: 639



The Food Acquisition and Purchase Survey (FoodAPS)





The Household Food Security Survey Module

Publication Coverage by Source for Household Food Security Survey Module Total Distinct DOIs: 1537



Rural-Urban Continuum Code

Publication Coverage by Source for Rural-Urban Continuum Code Total Distinct DOIs: 2431



3.2 Journal Coverage

Now that we have compared journal coverage across the two citation databases, we next examine the publications within journals that are indexed in both Scopus and OpenAlex. We report these results for the full-text search approach and the seed-corpus approach in OpenAlex. ARMS Financial and Crop Production Practices The Census of Agriculture Food Access Research Atlas The Food Acquisition and Purchase Survey (FoodAPS) The Household Food Security Survey Module Rural-Urban Continuum Code 3.3 Publication Topics ARMS Financial and Crop Production Practices The Census of Agriculture Food Access Research Atlas The Food Acquisition and Purchase Survey (FoodAPS) The Household Food Security Survey Module Rural-Urban Continuum Code

3.4 Author and 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, with particular attention to coverage of underrepresented and Minority-Serving Institutions (MSIs).

To create a harmonsized dataset of institutional coverage across datasets, institutional affiliation data associated with each publication's athor(s) are linked to institutional records using IPEDS identifiers. Linking the publication metadata with IPEDS institutional data adds information not available in the publication affiliation data alone. This additional information includes public or private institution (control), degree level, MSI designation, and geographic location. Special attention is given to coverage of underrepresented institutions and Minority-Serving Institutions (MSIs).

3.4.1 IPEDS and MSI Classifications

First, a clean, panel-form dataset of U.S. higher education institutions, including consistent MSI designations over time is created. Two sources were used: (1) the MSI Data Project (Nguyen et al., 2023) identifies MSI institutions from 2017–2021 and (2) Rutgers CMSI identifies MSI institutions for 2022–2023. These datasets were cleaned and merged with IPEDS institutional data, filtered to include only 2- and 4-year institutions in the 50 U.S. states. Data cleaning steps included: addressing inconsistencies in eligibility labels, removing duplicates, and creating summary measures of MSI eligibility by year. The resulting visualization highlights both the number and percent of institutions designated as MSIs over time, with a sharp increase observed in 2022. The accompanying plot and source code are available in the report appendices⁷ and MSI⁸.



⁷IPEDS appendix available here

⁸MSI appendix available here

3.4.2 Results of Institutional Comparision

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

Topic		First Run	OpenAlex Total
ID	Topic Name	Count	Count
T11610	Impact of Food Insecurity on Health Outcomes	549	78661
T10010	Global Trends in Obesity and Overweight	272	111686
	Research		
T11066	Comparative Analysis of Organic Agricultural	247	41275
	Practices		
T12253	Urban Agriculture and Community Development	222	27383
T10367	Agricultural Innovation and Livelihood	186	49818
	Diversification		
T11464	Impact of Homelessness on Health and	175	101019
	Well-being		
T12033	European Agricultural Policy and Reform	137	88980
T10841	Discrete Choice Models in Economics and	126	66757
	Health Care		
T10596	Maternal and Child Nutrition in Developing	116	118727
-	Countries		
T11898	Impacts of Food Prices on Consumption and	113	29110
T110F0	Poverty	100	15000
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
111886	Risk Management and Vulnerability in	73	44755
T10996	Agriculture	71	04104
110226	Global Analysis of Ecosystem Services and Land	(1	84104
T10966	Use Bala of Maditamanaan Diat in Haalth Outcomer	70	76904
T10000	Optimal Operation of Water Resources Systems	70	70094 07570
T 10909	Hydrological Modeling and Water Resources Systems	70 60	97070 120016
110990	Management	09	132210
T11753	Forest Management and Policy	60	75196
T12008	Bural development and sustainability	54	69114
T12098 T10111	Remote Sensing in Vegetation Monitoring and	59 59	56452
110111	Phenology	04	00402
T10556	Global Cancer Incidence and Mortality Patterns	<u>4</u> 9	64063
T11711	Impacts of COVID-19 on Clobal Economy and	49 49	69059
	Markets	10	00000

Table 3: Top 25 Topics by First Run Count

Topic		First Run	OpenAlex Total
ID	Topic Name	Count	Count
T12724	Integrated Management of Water, Energy, and Food Resources	47	40148

Journal		First Run	OpenAlex
ID	Journal Name	Count	Total Count
S276462809	6Journal of Agriculture Food Systems and	57	825
	Community Development		
S115427279	Public Health Nutrition	51	3282
S206696595	Journal of Nutrition Education and Behavior	41	3509
S15239247	International Journal of Environmental Research	39	59130
	and Public Health		
S421020186	1Applied Economic Perspectives and Policy	39	647
S10134376	Sustainability	35	87533
S5832799	Journal of Soil and Water Conservation	34	556
S273939355	5Journal 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
S276459330	0Agricultural and Resource Economics Review	25	247
S421021215	7Frontiers in Sustainable Food Systems	23	3776
S63571384	Food Policy	20	1069
S69340840	The Journal of Rural Health	20	749
S421023482	4EDIS	18	3714
S19383905	Agricultural Finance Review	18	327
S119228529	Journal of Hunger & Environmental Nutrition	17	467
S43295729	Remote Sensing	14	33899
S273839706	8Land	14	9774
S80485027	Land Use Policy	14	4559

Table 4: Top 25 Journals by First Run Count

Author ID	Author Name	First Run Count	OpenAlex 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

Table 5: Top 25 Authors by First Run Count Table