Methodology for Comparing Citation Database Coverage of Dataset Usage

Findings

2025-04-28

Table of contents

1 Overview

[Download PDF Version](#)

This report compares the differences between the Scopus and OpenAlex citation databases in their tracking of dataset mentions. The Census of Agriculture, produced by the USDA National Agricultural Statistical Services (NASS), provides information on U.S. farming operations, including production practices and land use. This dataset is used as a case study in this report to systematically compare the coverage, overlap, and differences in publications indexed by the two citation databases. The Census of Agriculture (also referred to as the "Ag Census") is selected for its frequent use in agricultural and economic research, making it an ideal dataset for assessing differences in publication coverage between Scopus and OpenAlex.

Secondary Objective: Institutional Representation

In addition to examining dataset mention coverage, the report also evaluates differences in institutional representation across Scopus and OpenAlex. Publication affiliation data are linked to institutional records using IPEDS identifiers to create a harmonized dataset across sources. This allows for assessment of how each citation database represents institutions based on characteristics such as control (public or private), degree level, MSI designation, and geographic location. Special attention is given to coverage of underrepresented institutions and Minority-Serving Institutions (MSIs). This analysis informs broader conversations about equity, transparency, and accountability in research data infrastructure.

2 Data Collection

To compare coverage across the two citation databases, publications that mention the Ag Census must first be identified. The methods used to identify dataset mentions in Scopus and OpenAlex are described below.

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.

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.

2.2.1 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.[1]

To implement these searches efficiently, the OpenAlex API was accessed using the `pyalex` Python package.[2]

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 the Census of Agriculture, the query 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", "book-chapter") - `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

---

[1] 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.

[2] `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.

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 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 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 mentions of the Census of Agriculture and related USDA datasets.

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

- Language: English
- Publication Year: Post-2017
- Publication Type: Articles and reviews
- Open Access Status: Open-access publications only

Filtering was further refined by selecting publications associated with high-relevance topics (Table 1) lists the top 25 topics), top publishing journals (Table 2) lists the top 25 journals), and U.S.-affiliated authors (Table 3) lists the top 25 authors).

Each table presents two key columns to help interpret the selection process. 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.

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 tradeoff supports a practical balance between depth and feasibility in building the final dataset.

Table 1: Top 25 Topics by First Run Count

Topic ID	Topic Name	First Run Count	OpenAlex 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

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

Table 2: Top 25 Journals by First Run Count

Journal ID	Journal Name	First Run Count	OpenAlex 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

Journal ID	Journal Name	First Run Count	OpenAlex Total Count
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 3: Top 25 Authors by First Run Count Table

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

Author ID	Author Name	First Run Count	OpenAlex Total Count
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

The resulting seed corpus contained approximately 1,774,245 (**CHECK WITH RAFAEL**) unique publications. An initial download of full-texts achieved a success rate of approximately 35%, corresponding to an estimated 625,000 accessible full texts. Local full-text searches were planned using this subset to improve detection of dataset mentions beyond what was possible through OpenAlex's built-in search functions.

While the seed corpus approach allowed 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.

3 Results Comparing Publication and Journal Coverage

An objective of this report is to understand differences in publication coverage across Scopus and OpenAlex. Specifically, we ask: (1) how many and which publications referencing the Ag Census appear in each citation database, and (2) how many and which journals publishing these articles overlap between the two sources. In addition, this analysis aims to determine whether the different search approaches used to find dataset mentions in OpenAlex publications—the full-text search versus a seed-corpus approach—produce substantially different coverage.

This section presents two sets of results. In both cases, a seed-corpus approach is used to identify Ag Census publications in Scopus. We then compare these results to (i) publications identified through a full-text search in OpenAlex and (ii) publications identified using a seed-corpus approach in OpenAlex.

```r
ui <- fluidPage(
  plotOutput("venn_plot")
)

server <- function(input, output, session) {

  output$venn_plot <- renderPlot({

    fit <- euler(c(
      "Scopus" = 4207 + 405 + 263 + 100,        # Only Scopus + estimated triple overlap
      "FullText" = 761 + 405 + 5 + 100,         # OpenAlex Full Text + estimated triple overl
      "SeedCorpus" = 940 + 263 + 5 + 100,       # OpenAlex Seed Corpus + estimated triple ove
      "Scopus&FullText" = 405 + 100,             # Scopus & FullText total
      "Scopus&SeedCorpus" = 263 + 100,           # Scopus & SeedCorpus total
      "FullText&SeedCorpus" = 5 + 100,           # FullText & SeedCorpus total
      "Scopus&FullText&SeedCorpus" = 100         # assumed triple overlap
    ))

    plot(fit,
         fills = list(fill = c("orange", "lightblue", "lightblue"), alpha = 0.6),
         edges = TRUE,
         labels = list(font = 4),
         quantities = TRUE,
         lty = c("solid", "dashed", "dashed"),   # Line type: Scopus solid, OpenAlex dotted
         col = c("black", "black", "black")        # Border color: all black
    )
  })
}

shinyApp(ui, server)
```
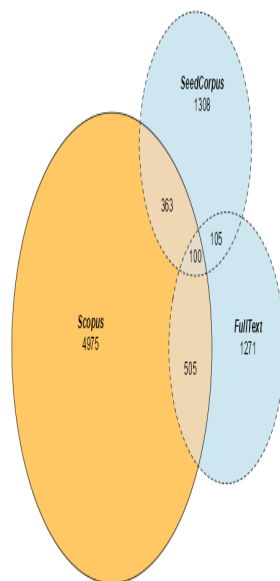
Listening on http://127.0.0.1:5361

Figure 1: Overlap of Scopus (paywall) and OpenAlex Full Text and Seed Corpus (open access) publications.

Set Combination	Meaning	Count
Scopus	Only Scopus + overlaps	4975
Full Text	Only OpenAlex Full Text + overlaps	1271
Seed Corpus	Only OpenAlex Seed Corpus + overlaps	1308
Scopus Full Text	Publications in Scopus and Full Text (including triple overlap)	505
Scopus Seed Corpus	Publications in Scopus and Seed Corpus (including triple overlap)	363
Full Text Seed Corpus	Publications in Full Text and Seed Corpus (including triple overlap)	105
Scopus Full Text Seed Corpus	Publications in all three sources	100

3.1 Method 1 (Full Text Search)

The first comparison looks at publication and journal coverage between the Scopus (using the seed-corpus method) and OpenAlex (using the full-text search). The results describe how many publications mention the Ag Census in each database and evaluates the degree of overlap between them.

Table 1 displays the coverage of publications that referenced the Ag Census data. The table distinguishes between results that appear in both databases (column 1), only in OpenAlex (column 2), and only in Scopus (column 3). Column 4 is a total of all distinct publications or journals across both databases. The first row reports the number of individual publications found in each category. The unit of analysis is at the publication level. The second row reports the number of journals that include the Ag Census publications, again broken out by their appearance in one or both datasets. The unit of analysis is at the journal level.

ADD TABLE 1 HERE

The main takeaway from Table 1 is that there is little overlap in publications and journals between the two databases when using OpenAlex's full-text search. According to this table, there are 5,473 unique publications referencing the Ag Census across both citation databases, appearing in 2,686 unique journals, identified by their ISSNs. The number of overlapping publications is 505 (9.23%), with the majority of publications referencing Ag Census are picked up only by Scopus (4,207 or 76.87%), and a smaller share is identified only in OpenAlex (761 or 13.9%).

Journal coverage shows a similar pattern. Of the journals that include at least one Ag Census publication, 247 (9.2%) are shared between the two databases, 2,362 (87.9%) are found only in Scopus, 77 (2.9%) only in OpenAlex.

These results show that the coverage of publications mentioning the Ag Census and the journals in which they are found is much more extensive in Scopus, and suggest that OpenAlex's full-text search along may miss many dataset mentions.

3.2 Method 2 (Seed Corpus Approach)

Based on the pattern observed from the full-text search in OpenAlex, the differences likely arise, at least in part, from limitations in how OpenAlex processes and indexes text. Specifically, we found that OpenAlex's full-text search does not index references as searchable text—they are stored as pointers, not included in the searchable body of the publication. In addition, n-gram-level metadata that might capture mentions of the Ag Census outside the main text was not accessible for the full set of publications. To address these limitations and create a more consistent comparison with Scopus, we applied a seed-corpus approach to OpenAlex, targeting a curated set of authors, journals, and topics associated with Ag Census use.

While this method can overcome the limitations of the OpenAlex full-text search, it is computationally more intensive. To limit the cost of the search corpus method, a list of the top authors, topics, and journals is provided, as described above. This list serves as a set of filters through which the search corpus is applied. Mentions of the Ag Census are then searched for within publications meeting these criteria.

ADD TABLE 2 HERE

Next, we compare the degree of overlap across the different search methods, focusing on OpenAlex publications that also overlap with Scopus records. Specifically, we examine whether the OpenAlex full-text search and the OpenAlex seed-corpus approach identified the same publications and journals referencing the Census of Agriculture, or whether each method uncovered distinct sets of works.

Table 3 summarizes the overlap between publications identified through the two OpenAlex search methods, restricted to publications that also appear in Scopus. As reported earlier, 505 publications were identified in both Scopus and OpenAlex using the OpenAlex full-text search (Table 1), while 363 publications were identified using the OpenAlex seed-corpus approach (Table 2). Table 3 specifically shows how many of the 505 full-text search publications were also found in the seed-corpus set. The comparison reveals that 105 publications are shared between the two methods, representing 20.8% of the full-text set and 28.9% of the seed-corpus set. These results suggest that the choice of search strategy meaningfully influences which Scopus-linked publications are recovered in OpenAlex.

ADD TABLE 3 HERE

Table 4 summarizes the overlap between journals identified through the two OpenAlex search methods, restricted to journals linked to publications that also appear in Scopus. As noted previously, the OpenAlex full-text search and seed-corpus approach each identified a set of journals containing publications referencing the Census of Agriculture. Table 4 specifically

shows how many journals were common to both sets. A total of 137 journals were shared between the two methods, representing 55.5% of the full-text set and 49.6% of the seed-corpus set. Among these shared journals, 19 were part of the original list of top journals used to construct the seed corpus.

ADD TABLE 4 HERE

4 Results Comparing Publication Coverage Disaggregated by Journal

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.

4.1 Method 1 (Full-Text Search)

Table 5 provides journal-level detail on the overlap between publications indexed in Scopus and OpenAlex, based on the full-text search results in OpenAlex. Each row corresponds to a journal included in the analysis. The set of journals included here matches the group of overlapping journals reported in Table 1—that is, journals where publications were found in both Scopus and OpenAlex.

The table is divided into two sections: overlap statistics and publication counts. The overlap statistics report three measures. The percentage labeled "% Both" indicates the share of a journal's publications that were found in both Scopus and OpenAlex. "% Scopus" shows the share of publications that appeared only in Scopus, while "% OpenAlex" shows the share of publications that appeared only in OpenAlex. Together, these columns summarize how consistently each journal's publications are represented across the two databases.

The publication counts section reports the number of overlapping and non-overlapping publications for each journal. "Pubs Both" shows the number of publications found in both Scopus and OpenAlex, while "Pubs Scopus" and "Pubs OpenAlex" show the number of publications found exclusively in one database. The final column, "Total Pubs," provides the total number of distinct publications associated with each journal across both databases.

Finally, the table also includes each journal's 2022 Scopus CiteScore to provide additional context on journal prominence.

Reading across a row, the table allows for direct comparison of database coverage at the journal level, highlighting journals where coverage between Scopus and OpenAlex aligns closely and those where substantial gaps exist.

ADD TABLE 5 HERE

To further characterize database coverage, we calculated weighted average CiteScores for the journals analyzed in Table 5. The weighted averages account for both the number of publications associated with each journal and the journal's 2022 Scopus CiteScore. For Scopus, the weighted average CiteScore, calculated by weighting each journal's CiteScore by its number of publications found in Scopus (including shared publications), is approximately 12.99. For OpenAlex, the weighted average CiteScore, using the number of publications found in OpenAlex (including shared publications), is slightly higher at 13.23.

These results suggest that, among the overlapping journals, OpenAlex tends to recover a slightly more citation-intensive set of publications compared to Scopus. However, the difference in weighted averages is modest, indicating broadly similar coverage of higher-impact journals across the two databases. To formally test the difference in weighted average CiteScores between Scopus and OpenAlex, we calculated a t-statistic for the difference in means. The resulting t-statistic is approximately -0.15, indicating that the difference between the two weighted averages is not statistically significant. This suggests that, within the set of overlapping journals analyzed, Scopus and OpenAlex recover a similarly citation-intensive set of publications.

4.2 Method 2 (Seed Corpus Approach)

We next examine the results obtained using the OpenAlex seed corpus approach.

Table 6 summarizes publication-level overlap statistics for journals identified using the seed corpus approach. As with Table 5, the table is restricted to journals associated with publications that overlap between Scopus and OpenAlex. The overlap statistics show the percentage of each journal's publications found in both databases (% Both), found only in Scopus (% Scopus), or found only in OpenAlex (% OpenAlex). Publication counts are also reported separately for shared and database-specific publications, along with the total number of publications associated with each journal. Each journal's 2022 Scopus CiteScore is included to provide additional context on journal prominence.

ADD TABLE 6 HERE

To further compare database coverage under the seed corpus approach, we calculated weighted average CiteScores for Scopus and OpenAlex. The weighted averages account for both the number of publications associated with each journal and the journal's 2022 Scopus CiteScore. For Scopus, the weighted average CiteScore is approximately 11.90, while for OpenAlex it is higher at approximately 14.10.

A t-statistic for the difference in means was also calculated to assess whether the difference is statistically meaningful. The resulting t-statistic is approximately -2.18, indicating that the difference is statistically significant. These results suggest that, when using the seed

corpus approach, OpenAlex recovers a more citation-intensive set of publications compared to Scopus.

5 Institutional Comparison

Each of the featured citation databases represent some portion of the global research landscape, yet their inclusion criteria and institutional coverage vary in ways that may inform disparities. Our goal is to assess which institutions are represented in each source, with particular attention to coverage of underrepresented and Minority-Serving Institutions (MSIs). By building a harmonized dataset that links IPEDS identifiers to institutional records across Scopus, OpenAlex, and Dimensions, we aim to evaluate how these citation databases include or exclude institutions across institutional characteristics such as control, level, MSI status, and geography. This analysis informs a broader conversation about equity, transparency, and accountability in research data systems.

Step 1: Clean IPEDS and MSI data

In this section, I document the construction and visualization of MSI (Minority-Serving Institution) eligibility trends from 2017 to 2023, as part of the broader effort to compare institutional coverage across Scopus, OpenAlex, and Dimensions. To support that analysis, we needed a clean, panel-form dataset of U.S. higher education institutions, including consistent MSI designations over time. I compiled and cleaned data from multiple sources—the MSI Data Project (Nguyen et al., 2023) for 2017–2021 and Rutgers CMSI lists for 2022–2023—and merged these with IPEDS institutional data, filtered to include only 2- and 4-year institutions in the 50 U.S. states. I addressed inconsistencies in eligibility labels, resolved duplicates, and created 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 included for transparency. All additional details are available in the IPEDS[3] and MSI[4] Appendices.

---

[3]IPEDS appendix available here

[4]MSI appendix available here

MSI Eligibility Over Time

Jump from 2021 to 2022: 400
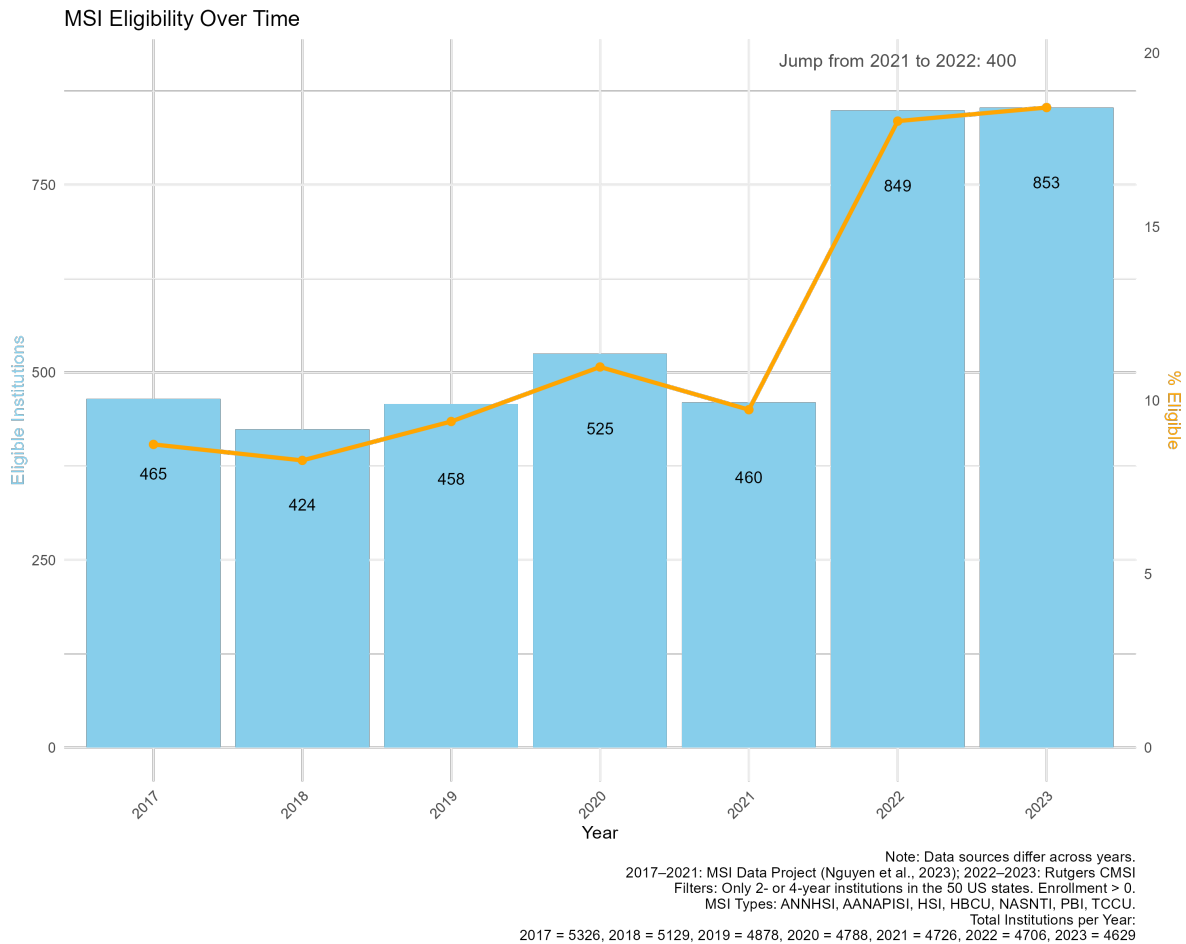
6 Summary of Findings

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).

Publication-level comparisons disaggregated by journal showed that, under the full-text search approach, OpenAlex and Scopus recovered broadly similar sets of citation-intensive publications. The weighted average CiteScores were 13.23 for OpenAlex and 12.99 for Scopus, with a t-statistic of -0.15, indicating no statistically significant difference. In contrast, under the seed corpus approach, OpenAlex recovered a more citation-intensive set of publications, with a weighted average CiteScore of 14.10 compared to 11.90 for Scopus. The t-statistic of -2.18 indicates that this difference is statistically significant.

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.