Methodology for Comparing Citation Database Coverage of Dataset Usage

Findings

2025-04-26

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.

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