Publication status: Preprint has been published in a journal as an article
DOI of the published article: https://doi.org/10.1162/qss.a.17

On the Open Road to Universal Indexing: OpenAlex and Open Journal Systems

Diego Chavarro, Juan Pablo Alperin, John Willinsky

https://doi.org/10.1590/SciELOPreprints.11205

Submitted on: 2025-02-04
Posted on: 2025-02-04 (version 1)
(YYYY-MM-DD)

On the Open Road to Universal Indexing: OpenAlex and Open Journal Systems

Diego Chavarro, Juan Pablo Alperín**, John Willinsky****

** School of Publishing, Simon Fraser University, Burnaby, BC, Canada (ORCID https://orcid.org/0000-0001-9116-0891), corresponding author (diego_chavarro@sfu.ca)*
*** School of Publishing, Simon Fraser University, Burnaby, BC, Canada (ORCID https://orcid.org/0000-0002-9344-7439)**, corresponding author (juan@alperin.ca)*
**** Graduate School of Education, Stanford University, Stanford, California, USA (ORCID https://orcid.org/0000-0001-6192-8687)*

31st of January 2025

Abstract

This study examines OpenAlex's indexing of journals using Open Journal Systems (JUOJS), reflecting two open source software initiatives supporting inclusive scholarly participation. By analyzing a dataset of 47,625 active JUOJS, we reveal that 71% of these journals have at least one article indexed in OpenAlex. Our findings underscore the central role of Crossref DOIs in achieving indexing, with 97% of the journals using Crossref DOIs included in OpenAlex. However, this technical dependency reflects broader structural inequities, as resource-limited journals, particularly those from low-income countries (47% of JUOJS) and non-English language journals (55%-64% of JUOJS), remain underrepresented. Our work highlights the theoretical implications of scholarly infrastructure dependencies and their role in perpetuating systemic disparities in global knowledge visibility. We argue that even inclusive bibliographic databases like OpenAlex must actively address financial, infrastructural, and linguistic barriers to foster equitable indexing on a global scale. By conceptualizing the relationship between indexing mechanisms, persistent identifiers, and structural inequities, this study provides a critical lens for rethinking the dynamics of universal indexing and its realization in a global, multilingual scholarly ecosystem.

Keywords

Open Journal Systems (OJS); OpenAlex; Scholarly Indexing; Digital Object Identifiers (DOIs); Global Scholarly Visibility; Equitable Knowledge Representation

1. Introduction

In a recent study we attempted to “recalibrate” the scope of scholarly publishing by describing a dataset of over 25,000 journals that publish their contents using Open Journal Systems (OJS), an open source publishing platform developed by the Public Knowledge Project (PKP), with which the authors of this paper are affiliated (Khanna et al., 2022). Identifying this many “journals using OJS” (JUOJS), we argued, significantly challenges estimates placing the number of scholarly journals worldwide at around 30,000 (e.g., Altbach & de Wit, 2018). In the intervening years, we have uncovered more than 20,000 additional JUOJS, further suggesting that, not only does an estimate of 30,000 “fall woefully short” (Khanna et al., 2022; p. 927), it continues to be eclipsed by a rapidly growing volume of scholarly activity worldwide. While recognizing that the JUOJS only share a common open source platform, we argued then, as we do now, that these journals offer an opportunity to observe aspects of scholarly publishing activity that have largely been overlooked in previous analyses.

In fact, one of the findings from our earlier study was that the need for recalibration stems from a significant number of journals being absent from major bibliographic indexes and databases. In particular, we documented how in 2022, some of the most widely used databases for analyzing scholarly publishing, Web of Science (WoS) and Scopus, contained only 1% and 7% of the 25,671 JUOJS publishing in 2020, respectively. In contrast, we found that JUOJS were significantly better represented in newer, more inclusive, databases, with 55% of JUOJS found in Dimensions and 64% in OpenAlex. The presence of JUOJS in these indexes can serve as an indicator, albeit an imperfect one, of the extent to which indexing sources are facilitating knowledge discovery on a global scale in the service of research and scholarship.

In the present study, we extend the analysis of the indexing of JUOJS to the 47,625 journals identified most recently (Khanna et al., 2024) with ISSNs verified through the ISSN International Center. In doing so, we seek to go further than simply ascertaining the inclusion of JUOJS in any bibliographic index, and instead answer questions about the mechanisms and infrastructures available to these journals, and the editorial practices that determine their use, that may be leading to their inclusion or omission. Ultimately, we seek to answer the question: *what are the characteristics of journals that lead to their being indexed in open bibliographic databases and how these characteristics reflect structural disparities in global scholarship?* Given the importance of persistent identifiers, we are especially interested in understanding the reliance on Digital Object Identifiers (DOIs) for journals to be found and indexed by databases. Our analysis focuses on JUOJS and uses these journals as a case study through which to understand the relationship between characteristics, such as their country of publication, discipline, languages, editorial characteristics and web visibility and authority, and access to Crossref DOIs and OpenAlex—two open scholarly infrastructure services that aim to provide inclusive (i.e., non-selective) indexing.

2. Literature review

One of the motivations to build journal indexing databases was an ambitious vision of covering all the scientific literature generated worldwide, so that indexing could offer access to universal knowledge (Chavarro 2017, p. 56-58). The goal of indexing is to facilitate researchers' knowledge discovery, so that, in relation to the most relevant work that has come before, they can build on, augment, critique, and set new directions. In Robert Merton's (1942) classic statement of the scientific ethos, indexing can be seen to ideally serve three of his four "institutional imperatives," including science's communal interests given to the widest possible sharing of research; its universalist ideals in which everyone may participate and contribute to science; and its organized skepticism, which calls for the close scrutiny of related studies for their empirical adequacy.

In his seminal paper, Bradford (1934,1985) examined the distribution of scientific papers in the subject of "applied geophysics and lubrication" across journals and uncovered the inefficiencies in indexing literature. He identified three groups of journals based on their output, finding that many relevant papers were published in lesser-known journals thus making it difficult to find that knowledge, leading to his formulation of "Bradford's law of scattering."[1] While Bradford aimed to promote inclusion by highlighting overlooked journals, his findings were paradoxically utilized by librarians to determine a "core" set of journals to purchase, often justifying the exclusion of many others. The rising costs of journals in the 20th century further pressured libraries to limit their acquisitions, motivating studies to optimize journal selection for user satisfaction. This ironic application of Bradford's work laid the groundwork for Eugene Garfield's "law of concentration," (Garfield 1971) which strengthened the concept of core journals in library acquisitions.

The notion of core journals has played a significant role in the development of bibliometrics. The creation of Scopus in the mid-2000s, while expanding the indexing of WoS (Mongeon & Paul-Hus, 2016), further entrenched a perception of knowledge exclusivity, selectively highlighting a curated list of "quality" journals (Guédon 2001; Chavarro 2017, p. 58-59). This transition not only influenced researchers' perceptions of value, but also (inadvertently) marginalized part of the global scientific community, who communicate in languages other than English, produced knowledge in disciplines considered to be of lesser scientific value, or whose publishing houses reside in countries considered marginal (Chavarro, Ràfols, & Tang 2018).

However, now that virtually all journal articles are published online, and the technologies exist to make it possible to comprehensively index them, the goal of universal indexing—or at least more inclusive indexing—has been reinvigorated. The goal first materialized in the mid-2000s when Google Scholar came onto the scene with a radically different approach. Multiple studies have examined the extent of the coverage of Google Scholar, as compared

[1] This law suggests that journals can be categorized into a core and several zones based on their article productivity, represented by a ratio of $1:n:n^2$, where 1 represents the core, n represents the second group with the same number of papers as the core but are less focused on the subject, and n^2 represent the third group who also publish the same number of papers as the core but are even less focused on the subject.

to WoS and Scopus, finding that Google Scholar indexed up to 300 Million records by 2019 (Delgado, Orduña-Malea & Martín-Martín 2019) and recovers up to 88% of citations to well-known papers, beating comparable citation databases on these two indicators (Martín-Martín et al. 2021). However, the lack of access to the underlying data has made Google Scholar a challenging source for bibliometrics. Microsoft developed a similar web-indexing approach that also covered more documents than Scopus and WoS (Wang et al. 2020; Martín-Martín et al. 2021). When Microsoft Academic Graph (MAG) was discontinued in 2022, the company released its complete data set for others to use. OurResearch, a nonprofit that develops open source software tools for research discovery and analysis, drew on the MAG data in creating OpenAlex as an index intended to be comprehensive and inclusive, while making its data open (Priem et al., 2022).

While university administrators and others are evaluating or adopting OpenAlex as a replacement for the closed and selective competitors (OurResearch 2024; CWTS 2024; Sorbonne, 2024), the bibliometrics community has been busy scrutinizing the metadata quality and completeness of the data source. OpenAlex has been found to index far more works than any other source (Culbert et al., 2024; Jiao et al., 2023; Alperin et al., 2024), but also highlight issues of completeness and correctness of metadata (e.g., Alperin et al., 2024; Delgado-Quirós & Ortega, 2024; Culbert et al., 2024; Mongeon et al., 2023; Akbaritabar et al., 2023; Jahn et al., 2023; Alonso-Alvarez & van Eck, 2024). While understanding the issues in any given database is important, these same issues can be used to better understand the underlying scholarly publishing infrastructures that have led to them in the first place, as well as leading to their improvement.

Parallel to the development of bibliographic databases has been the development and adoption of various publishing infrastructures, including the growth in preprint servers (Chtena et al. 2024), better federated access to institutional repositories (COAR 2024), growing use of the Directory of Open Access Journals (Hugar 2019), and, as mentioned at the outset, the use of OJS. These services and platforms have themselves been supported by the development and adoption of persistent identifiers, namely, DOIs (for documents), ORCIDs (for people), and, most recently, RORs (for institutions). Initiatives like OpenAlex have been able to make use of the way these infrastructures organize scholarly metadata, but this creates a potential dependency between access to certain infrastructures and inclusion in bibliographic databases.

Given OpenAlex's inclusive indexing goals and the availability of structured metadata offered by OJS, one might expect OpenAlex's complete coverage of OJS. However, our earlier study found that only 64% of JUOJS were covered at the time by OpenAlex (Khana et al., 2022). While this percentage is significantly higher than the JUOJS in WoS (1%) and Scopus (7%), it still lacks an important percentage of journals that fall outside the oligopoly found in mainstream databases (van Bellen, Alperin & Larivière, 2024).

This discrepancy stems from multiple challenges in scholarly indexing. OpenAlex faces limitations in comprehensive coverage due to the dynamic nature of publishing, with regional, and disciplinary biases persisting despite its more inclusive approach (Alperin et al., 2024; Cespedes et al., 2024; Maddi, Maisonobe, & Boukacem-Zeghmouri 2024). While offering more balanced representation of non-English languages compared to traditional databases (Cespedes et al., 2024), OpenAlex still struggles with metadata completeness and accuracy

for some of the expanded content (Zhang et al., 2024; Culbert et al., 2024). These factors underscore the ongoing need for improvements in bibliographic database inclusivity (Delgado-Quirós & Ortega, 2024).

This study therefore makes use of detailed publication data of JUOJS that we have collected as part of our work with PKP to examine the nature of the coverage of this subset of the scholarly literature. While we simultaneously collaborate with OpenAlex to improve the indexing of JUOJS, we take advantage of a snapshot of OpenAlex from 2024 to more closely examine the properties of those publications that have, until now, eluded indexing in this database. While the indexing issues of JUOJS may be addressed through direct cooperation (currently ongoing), we expect the findings of this study to assist the community in addressing challenges faced by other smaller publishers beyond those using OJS.

The reasons why any individual work may not be indexed in OpenAlex are several, but largely stems from their absence in OpenAlex's underlying sources—most significantly Crossref and the now defunct MAG (OpenAlex, n.d.). The reliance on DOIs has been noted by experienced editorial managers, such as Ansorge (2022), who in a discussion paper describes her experience with managing a journal that does not assign DOIs and compares its indexing to journals that do. She finds that OpenAlex, Dimensions, and other services do not index the journal's papers despite the journal having ample presence in its area.

This reliance on DOIs has led to a significant underestimation of bibliometric indicators such as paper count, authorship, h-index, citations, and other indicators (Khurana et al. 2022). At a conceptual level, Okune and Chan (2023) argue that the DOI system fosters a centralization of power, predominantly controlled by traditional publishers. This control effectively marginalizes smaller publishers, as the associated DOI fees—while seemingly affordable from the standpoint of high-income countries—present substantial obstacles for journals in low-income nations or for organizations with limited resources. Crossref its recognizes this risk, and attempts to mitigate it through tiered pricing and through a Global Equity Membership (GEM) program, which removes financial barriers for any journal publishing from low-income countries. Still, Okune and Chan (2023) highlight that the widespread adoption of DOIs has led to a misleading perception that they serve as a certification of quality for the research produced, whereas their main role is to ensure accuracy in locating the article and the tracking of citations to it.

In these ways, the DOI simultaneously serves as a door to universal indexing and as an obstacle to achieving this goal. It is, then, a key variable to be studied in coverage analyses, which have been traditionally focused on the disciplinary, linguistic, and geographical biases of databases (van Leeuwen et al. 2001; Sivertsen & larsen 2012; Chavarro, Ràfols, & Tang 2018; Khana et al. 2022). Along the same lines, we also include other less explored variables in coverage studies that are important for indexing in the era of digital publishing, namely related to visibility, discoverability and web authority. In this way we attempt to add to our understanding of the factors that may contribute to or hamper Bradford's goal of universal indexing.

3. Data and Methods

Data Sources

This study makes use of a variety of data sources: the PKP Beacon to obtain a list of active JUOJS and their publications; the registry kept at ISSN.org to validate the ISSNs of these journals; the DOI resolution services maintained by DOI.org and by Crossref to validate the DOIs identified for JUOJS; the database maintained by CommonCrawl to obtain the OpenPageRank indicator (Domcop 2024), which is a measure of a webpage reputation built on the CommonCrawl data (CommonCrawl Org 2024); the indices of both DOAJ and Scopus to identify their lists of indexed journals; the World Bank (World Bank 2024) for GDP per capita data; and OpenAlex (Priem, Piwowar & Orr 2022) to identify the coverage of the JUOJS. Below we describe how we used each of these data sources to build our variables.

Journals Using OJS (JUOJS)

This study relies on data gathered from the PKP Beacon, a feature introduced into OJS in 2015 that allows PKP to notify OJS users about security updates and software upgrades. The beacon also notifies PKP of the web location of each journal's metadata harvesting API, which we subsequently used to collect journal metadata, such as the journal name and ISSN, and article metadata, such as titles, abstracts, publication dates, and DOIs. As in our previous study, we considered a journal to be "active" in a given year if it published at least five documents that year (a threshold established by the Directory of Open Access Journal (DOAJ) in 2020). We began with the OJS journal list updated in 2024 (Khanna et al., 2024). This updated list included 50,920 journals with an ISSN, and that were deemed to be active from 2020 to 2023 (a 98% increase in the number reported for 2020; Khanna et al., 2022). For the present analysis, we included only the 47,625 journals that were deemed active for at least one year between 2020 and 2023, and whose registries could be validated against ISSN.org (see below). These journals were distributed among 17,447 installations (on average, each installation of OJS contains three journals). The journals in the data set averaged 15 items per year between 2020 and 2023, published 2,962,418 items during those years, and have published a total of 10.6 million items since inception.

ISSN Validation

As mentioned, to ensure we were working with active and valid journals, we compared the title, title language, and ISSN of each journal with the official ISSN registry by querying its search interface in JSON format. We compared the strings of these fields in OJS and the strings reported by ISSN.org for each print and e-ISSN available. As some journal titles are written in various alphabets, we first tried to identify the title language and then transliterated their titles to ASCII. To address potential variations in spelling and pronunciation that might occur across different entries of the same journal title, we employed the Soundex algorithm. This phonetic algorithm encodes words based on their pronunciation rather than their exact spelling, which was particularly useful because it allowed us to match journal titles that sound similar despite having minor spelling differences.

We then calculated different distance measures between the transliterated journal names and between the soundex codes: jaccard (Jaccard 1912; Choi, Cha & Tapert 2016), jaro-winkler (Winkler 1990), cosine (Salton & McGill 1983), levenshtein (Levenshtein 1966), hamming (Hamming 1950), Qgram (Ohkura et al. 2005) Additionally, we utilized term frequency / inverse document frequency (TF/IDF) (Sparck 1972) to identify journals with extended titles in OJS and key titles or acronyms in ISSN.org. We selected the measure that produced the shortest distance between journal titles, ranging from 0 (very different) to 1 (exact match) and considered valid those matches that passed a similarity threshold of 95%. We were able to validate the titles and ISSNs of 47,625 (94%) of the 52,920 active JUOJS.

Indexing

Using the validated ISSNs, we verified the presence of JUOJS in three indexing services: OpenAlex, the Directory of Open Access Journal (DOAJ) and Scopus. In all three cases, we considered a journal to be indexed by that service if either of the validated ISSNs (print or online) could be found in the API (OpenAlex), publicly available list (DOAJ) or in the list provided to subscribers (Scopus).[2]

Digital Object Identifiers (DOIs)

A DOI was present in the metadata for 2,196,697 articles published in the 47,625 active JUOJS with valid ISSNs, from 2020 to 2023. To verify these DOIs, we employed two methods: querying DOI.org and using the Crossref API, which is the most widely used registration agency for scholarly journals. This dual-verification process allowed us to cross-check the existence and validity of the DOIs.

Using DOI.org, we successfully validated 1,839,607 DOIs, representing approximately 84% of the total. Among these validated DOIs, 94% were registered with Crossref, 5% with DataCite, and 1% with other registration agencies. To ensure the accuracy of our DOI validation, we compared data from DOI.org (specifically for Crossref-registered DOIs) against Crossref.org. The results showed a 99.9% agreement between the two sources, confirming the reliability of our DOI search and validation process.

[2] OpenAlex does not index by journal, but rather, by works. As such, a journal's absence in this source does not necessarily mean that works from that journal are not present in the database.

PageRank

To have an indicator of a journal's visibility on the Web, we queried its Open PageRank in the CommonCrawl dataset (CommonCrawl Org 2024)), as has been done in previous bibliometric studies (e.g., Lin et al., 2023); Wang, 2022). CommonCrawl offers a free and open source global crawl of the Web that is updated monthly, and is frequently used in the training of machine learning and large language models (Facebook 2024). For each repository and journal URL, CommonCrawl returns an Open PageRank score between 0 (no page rank) and 10 (top page rank).

Gross Domestic Product

The OJS Beacon data identifies the journal's country through its ISSN, MARC record, DOAJ listing, and/or IP address. It was not able to identify a country for 463 journals. For all remaining journals, we queried the country code for each journal using the World Bank API for the country's Gross Domestic Product (GDP).

Calculated variables

We also constructed additional variables of journal and journal portal characteristics from the metadata present in the PKP Beacon.

Language of Publication

We ran Meta's Fasttext algorithm for language detection on each article title (Joulin et al. 2016). We then aggregated the number of papers by language and journal and classified each journal as being either monolingual (English), monolingual (non-English), multilingual (with English), multilingual (without English). Details of the definitions and the number of journals in each category can be found in Table 1.

Table 1. Number and percentage of journals by publication language

Language Category	**Category Description**	**Number of journals**	**Percentage**
Multilingual English	The articles in the journal are published in more than one language, including English, and none of them account for more than 90%.	26,281	55.2%
Monolingual English	More than 90% of the articles are published only in English	12,057	25.3%
Monolingual non-English	More than 90% of the articles are published in a single non-English language.	8,104	17.0%
Multilingual non-English	The articles in the journal are published in more than one non-English language, none of them account for more than 90%.	994	2.1%
N/A		189	0.4%

Total		47625	100.0%

Discipline

To identify the discipline of each journal, we trained a random forest algorithm on the multilingual embedding (Reimers 2019) of the journal titles using the list of DOAJ journals (DOAJ 2024). We used DOAJ as the training set because of its multilingualism, size, and global scope, and because DOAJ journal subject classifications have been manually curated by DOAJ staff. As such, our resulting classification of JUOJS uses the top-level Library of Congress Classification (LCC), as per the DOAJ's current practice. We manually verified a random set of 100 journals in the unseen data and found 78 correct classifications, in line with the model's accuracy (which was calculated at 80%, using a simple percentage of correctly predicted instances). The number of journals in each discipline can be found in Table 2.

Table 2. Disciplines of Journals Using OJS*

Discipline	**Number of journals**	**Percentage**
Social Sciences	9,832	20.6%
Medicine	8,465	17.8%
Science	5,979	12.6%
Technology	5,596	11.8%
Education	5,090	10.7%
Philosophy. Psychology. Religion	3,456	7.3%
Language and Literature	2,961	6.2%
Law	1,633	3.4%
Agriculture	1,379	2.9%
Geography. Anthropology. Recreation	1,256	2.6%
Fine Arts	663	1.4%
History (General) and history of Europe	537	1.1%
Political science	437	0.9%
Auxiliary sciences of history	104	0.2%

*Only disciplines with more than 100 journals shown

Journal and Installation Characteristics

Finally, we calculated several variables related to each journal and its OJS installation (which could include multiple journals). In particular, we calculated: the number of journals in the installation, the total number of documents published per journal, and the year of first publication.

3.2 Methods

In seeking to understand which JUOJS could be found in OpenAlex and which are assigning DOIs, we opted to use classification trees (Breiman 1983). Classification trees provide a visual representation of the main variables affecting coverage and offer a hierarchical relationship between variables that can be helpful for understanding their relative importance. They are especially suitable for the classification of categorical outputs based on categorical and continuous inputs, as is the case for our data and questions.

The classification tree structure consists of a root node for the entire population that is split into child nodes based on decision rules derived from the data, leading to further splits at subsequent levelsTo decide each split, classification trees use a criterion, such as the Gini Impurity Indicator or the Information Gain Indicator.. For our analysis, we used the Gini Impurity indicator to create mutually exclusive subgroups that are as homogeneous as possible (Hückstädt, 2023, p 1935) by quantifying the probability of incorrectly classifying a randomly chosen record.

In contrast with correlations and regression models, classification trees offer simple decision rules based on feature selection and interaction of the variables visualized in a relatively intuitive graphical output. Moreover, classification trees are suitable for use with untransformed variables, are able to model non-linear relationships (Chen, Wang, & Zhang 2011), and are also robust to outliers (Breinman, 2001, p. 10) and violations of homoscedasticity and normally distributed residuals (prerequisites for many regression analyses) (Hückstädt, 2023, p. 1937).

However, classification trees also face some challenges. One is that the tree solution is prone to overfitting (Kern et al., 2019, p. 75), which happens when the algorithm learns too well the structure of the observed data, leading to a bias that impedes generalization to new data. This is often managed by pruning the tree, thus avoiding tree structures that are too complex (Hückstädt, 2023). Also, the structures of the trees can be very sensitive to small variations in the data, which can guide to completely different structures based on such small variations (Strobl et al., 2008). To avoid these problems, we followed the recommended k-fold cross-validation (Chen, Wang, & Zhang 2011; Hückstädt, 2023). In this process, the dataset is divided into 'k' equal parts or folds. The model is then trained k times, each time using a different fold as the test set while the remaining k-1 folds serve as the training set. For example, in a 5-fold cross-validation, the data would be split into five parts, and the model would be trained and tested five times, each time using 80% of the data for training and 20% for testing, but with a different 20% segment each time. The results from all these iterations are then averaged to produce a more reliable estimate of the model's performance. Through this cross-validation process, a hyperparameter is produced that helps prevent overfitting and generates unbiased or less biased comparisons between the subtrees (Chen, Wang, and Zhang, 2011, p. 57)..

In our research, we fitted a first classification tree to find out the main variables and rules that increase the probability of a journal being indexed by OpenAlex. The variables used were: GDP per capita of the publishing country, number of journals in the publishing country, DOI registration agency (Crossref, DataCite, etc.), indexed in Scopus, indexed in DOAJ, total number of articles, number of journals in installation, earliest publication year, Open PageRank

of the journal, and OpenPage Rank of the installation. As the result showed an overwhelming importance of the Crossref DOI variable for determining whether a journal is indexing by OpenAlex, we fitted a second classification tree with the same variables (minus the DOI registration agencies) to better understand the characteristics that lead to a journal using Crossref DOIs..

In both trees all variables were fed into the model preserving their original values and scales, except for the DOI registration agencies variables which were dichotomized (1 if the journal had any DOIs registered with that agency, and 0 otherwise). To decrease the risk of overfitting and biasing the results, we performed a robustness check using nested k-fold cross-validation (Hückstädt, 2023). This technique consists of performing cross-validation within each loop of the standard cross-validation. The outer loop is used to estimate the generalization performance of the model, while the inner loop is used for hyperparameter tuning. This double validation reduces the overfitting of the model ensuring that each cross-validation assesses the accuracy of the model on data that it has not seen during training or hyperparameter tuning (Varma & Simon 2006). We used an outer validation of 5 and an inner validation of 3. We plotted the pruned trees to preserve simplicity and listed the variable relative importance based on mean decrease in accuracy (Hückstädt, 2023). Although we list the variables for information purposes, our interest remains in the main rules that are chosen by the tree, which may not include some of those variables.

Finally, we provided performance evaluation measures of accuracy (correct predictions over total predictions), precision (proportion of true positive predictions all positive predictions), recall (true positives over all actual positives), F1 score (the harmonic mean of precision and recall), and Area under the ROC curve (indicator of the trade-off between true positive rates and false positive rates at various threshold settings) (Fawcett 2006). We provide the percentage improvement of the classification accuracy over the baseline prevalence model using the Inter-model vigorish (IMV) measure (Domingue et al. 2021).

4. Results

4.1 Coverage in OpenAlex

OpenAlex has a significant coverage of JUOJS, including 33,654 (71%) of 47,625 journals This leaves 13,971 JUOJS uncovered by OpenAlex (Figure 1). Coverage by discipline resembles overall coverage, with 70% of coverage on average (SD = 0.03) (Figure 2). Similarly, coverage remains an average of 70% (SD=0.025) for high, upper-middle, and lower-middle income countries (Figure 3). However, low income countries have a 47% coverage, which is significantly lower than for the other income groups. In terms of languages, the average coverage is 60% (SD= 0.15), but two groups can be identified. Those including English (monolingual and multilingual), whose coverage is 74% and 71% respectively, and those not including English (monolingual = 64% and multilingual = 55%). Finally, Table 3 in columns 4 and 5 show that the majority of JUOJS have published less than 500 articles in total (89% of all active journals). As the number of total articles increases, the percentage of journals included in OpenAlex tends to increase.

Figure 1. Global coverage of JUOJS by OpenAlex

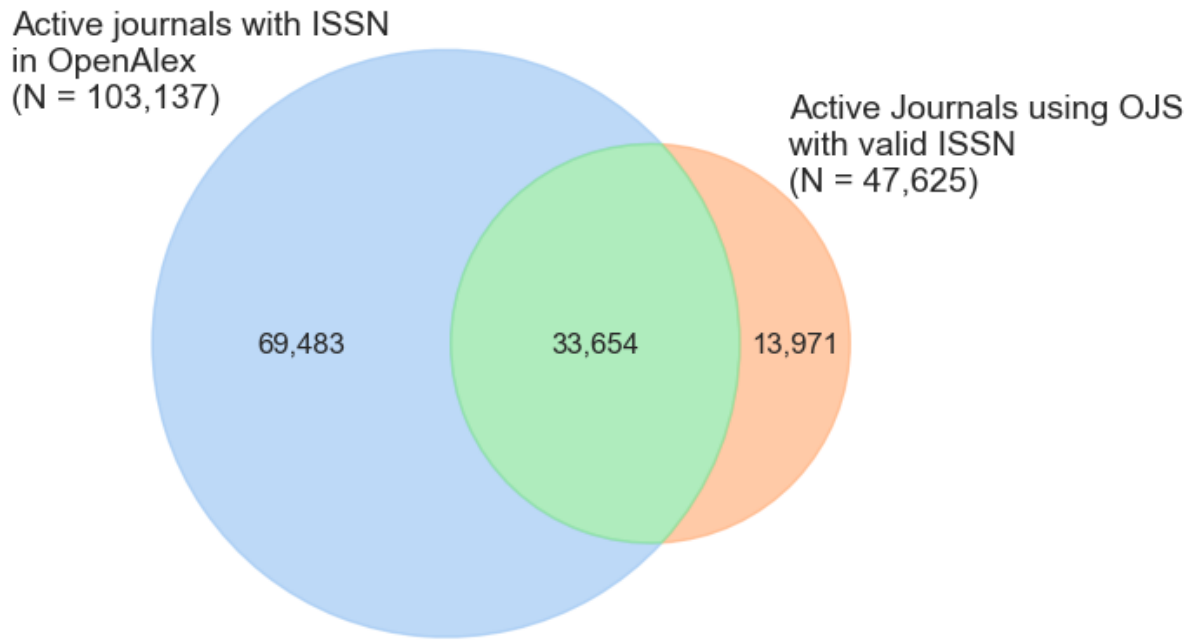
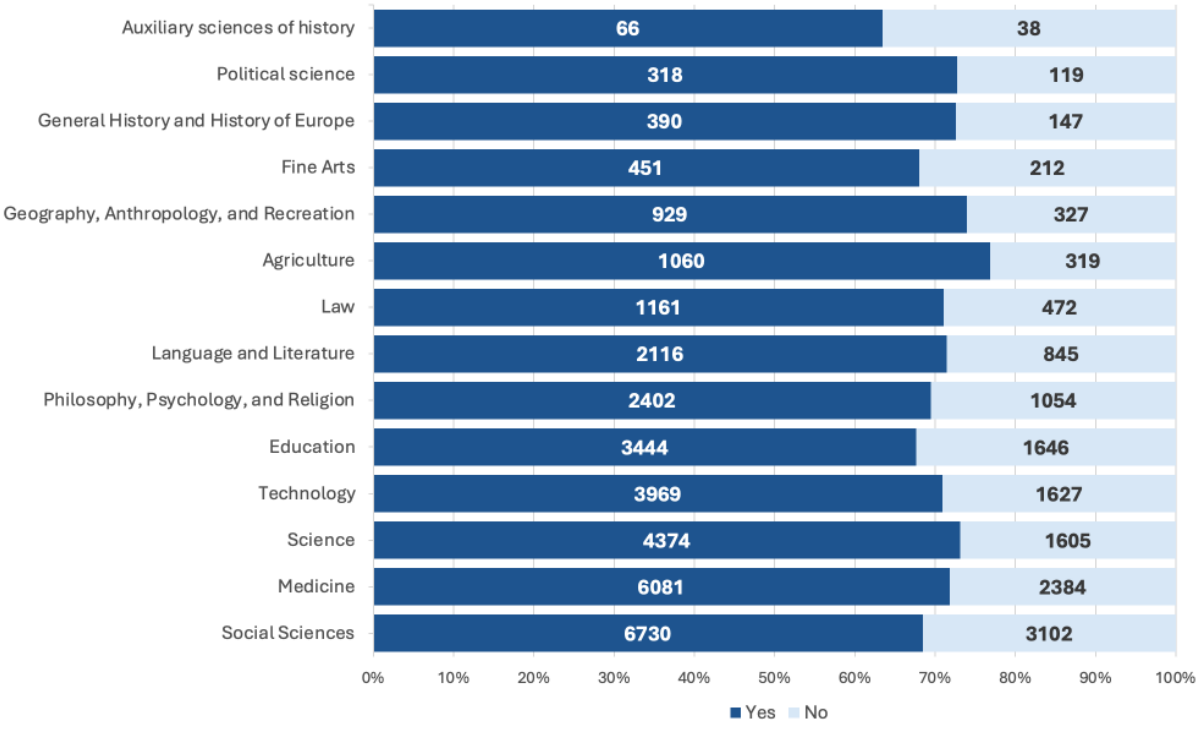
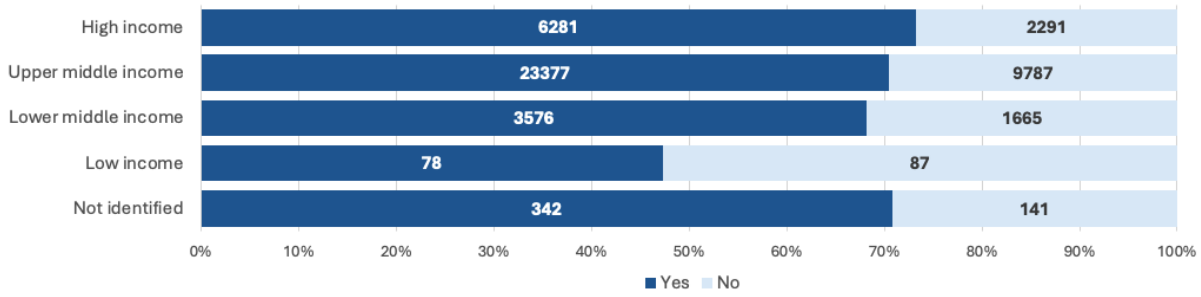


Figure 2. OpenAlex coverage of JUOJS by LLC discipline



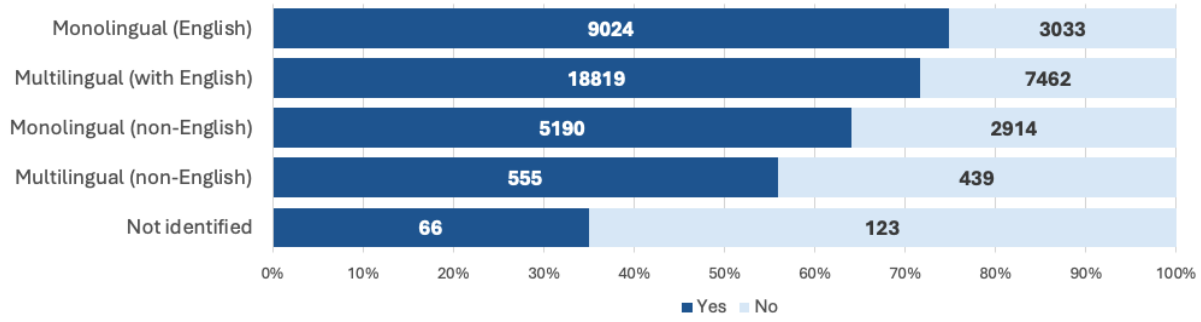
*Only disciplines with at least 100 journals are shown. Disciplines are ordered by total number of journals. Bars indicate percentages and numbers in the bars are the number of journals.

Figure 3. OpenAlex coverage of JUOJS by country income group*



*Income groups are defined by the World Bank (World Bank 2024).

Figure 4. OpenAlex coverage of JUOJS by language of the journal*



*Ordered by percent coverage

Table 3. Number and percentage of JUOJS covered and not covered by OpenAlex by journal's total article count*

Number of articles	In OpenAlex	Not in OpenAlex	Total	Percent in OpenAlex
<=100	15,461	8,918	24,379	63%
101 - 500	14,344	4,185	18,529	77%
501 - 1000	2,486	568	3,054	81%
1001 - 5000	1,288	274	1,562	82%
>5000	75	26	101	74%
Total general	**33,654**	**13,971**	**47,625**	**71%**

*Total number of articles are counted since first year of publication.

4.3 Factors influencing coverage in OpenAlex

We fitted a classification tree on the 47,625 active and validated JUOJS. The most relevant predictors are in Table 4. Appendix 1 provides the descriptive statistics table for the variables, grouped by inclusion in OpenAlex. The classification tree was performed using a nested cross-validation of 5 fold, which retains enough data in each training set to train the model effectively while still having a sufficient number of test sets to evaluate performance. We performed three folds for the inner cross-validation to allow for combinations within each outer fold and find the best hyperparameter for the classification tree. We provide the relative importance of each variable (Table 4) and a visual representation of the tree pruned at depth 3 (Figure 6). If a publication has a Crossref DOI, it has a 97% chance of being included. If it does not have a Crossref DOI but is listed in the Directory of Open Access Journals (DOAJ) and not in DataCite, the chance of coverage is 81%. Similarly, if it lacks a Crossref DOI and is not in DOAJ but is indexed in Scopus, it also has an 81% chance of being covered. If a publication has none of these identifiers (Crossref DOI, DOAJ, or Scopus), the likelihood drops to 41%. Additionally, if a publication does not have a Crossref DOI but is listed in DOAJ and DataCite, the chance of coverage is 34%. Besides the main rules, the relative importance of variables confirm the weight of Crossref DOIs in the model. Other variables were identified as important but did not provide a significant improvement in the final model's predictive power when compared to other variables.

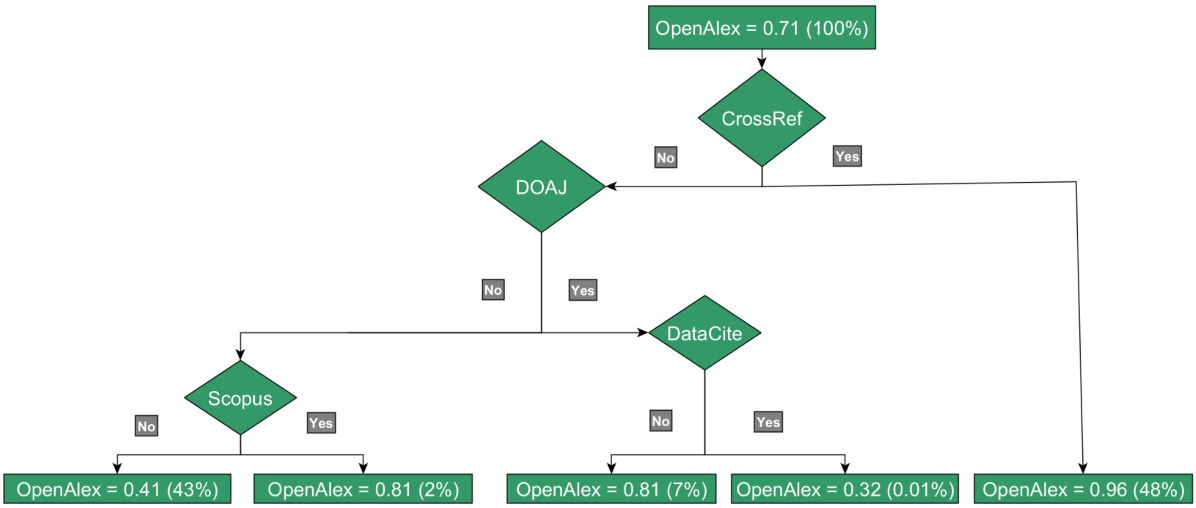
Table 4. Relative importance of variables in relation to being covered by OpenAlex*

Variable	Description	Relative Importance
Crossref DOI	has at least one DOI from Crossref	100.0
DOAJ	indexed in DOAJ	28.1
Total number of documents	total number of documents since inception	25.9
open Page Rank of the Endpoint	age rank of the End Point (1 - 10)	235.97
Scopus	Indexed in Scopus	15.9
Open Page Rank of the journal	Page rank of the journal	14.9
DataCite	has at least one DOI from DataCite	14.1
Monolingual-English	Journal has at least 90% of papers published exclusively in English	10.8

Repository Size	Number of journals in repository	10.7
earliest Year Publication	Year of publication of the first paper of the journal	7.8
GDP Per Capita	GDP per capita of the country of publication	5.9
Number of journals in the country	Number of JUOJS in the country of publication	2.9

*only variables greater than 1 shown

Figure 6. Decision tree of rules to being covered by OpenAlex*



*The tree model shows the rules for predicting coverage by OpenAlex and is based on the analysis of 47,625 journals. the numbers in each rectangle are the probability and percentage of journals (in brackets) meeting the criteria. The model achieved an accuracy of approximately 81%, which is 36% better (IMV = 0.36) than the prevalence baseline model of 0.71. The precision score suggests that 66% of the positive class predictions were accurate, which points to some challenges with false positives. On recall, the model effectively identified 72% of actual positive cases, reflecting a moderate capacity to capture relevant instances. The F1 score of 69% suggests a reasonable balance between precision and recall. Notably, the Area Under the Curve (AUC) reached 0.86, pointing to a strong ability to differentiate between positive and negative classes across various threshold settings.

4.3 Factors Influencing use of Crossref DOIs by JUOJS

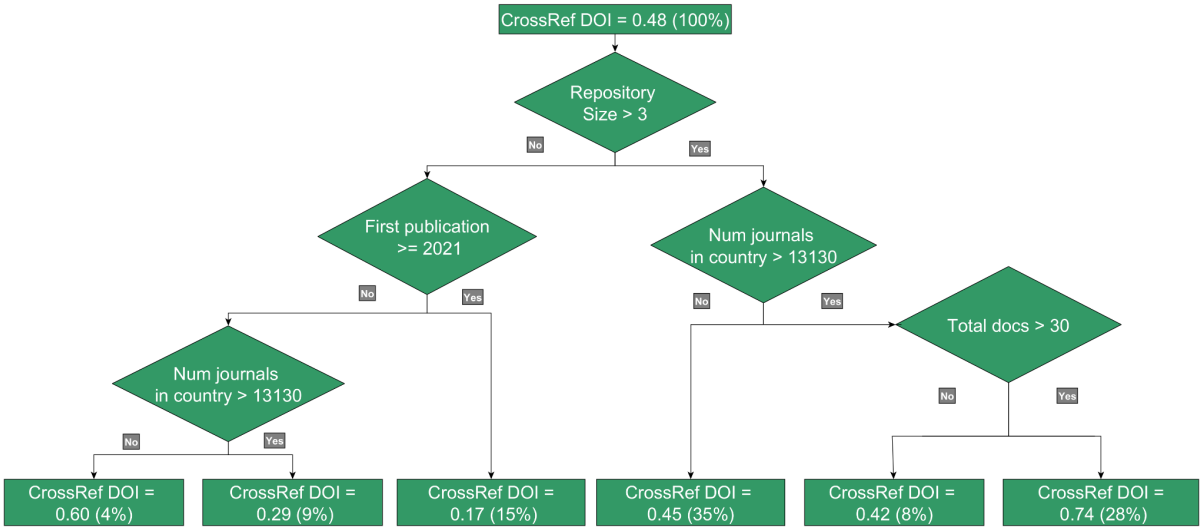
Given the reliance on Crossref DOIs for indexing in OpenAlex, we ran a second classification tree to identify the factors predicting whether JUOJS use Crossref DOIs. The main predictors resulting from the tree can be found in table 5. The classification tree was performed using the

same parameters as in the earlier classification tree (nested cross-validation of 5 fold for the outer cross-validation and 3 folds for the inner cross-validation). Descriptive statistics for the variables, grouped by usage of Crossref DOIs can be found in Appendix 2. The relative importance of each variable is in Table 5, and a visual representation of the tree pruned at depth 3 is in Figure 8. In this case, the classification tree rules indicate that journals have a probability of 0.73 if they are part of an installation with more than 3 journals, are published in a country with at least 13,130 JUOJS, and have published at least 30 documents. The next highest probability (0.59) occurs when the journal is part of an installation with less than 3 journals, the country has 13,130 or more JUOJS, and the earliest publication year is prior to 2021. Following this, a probability of 0.44 is predicted when the repository size is 3 or greater, but the country has fewer than 13,130 journals. A probability of 0.40 is found when the repository size is 3 or greater, the country has at least 13,130 journals, and the total number of documents is less than 30. The probability decreases to 0.29 when the repository size is less than 3, the country has fewer than 13,130 journals, and the earliest publication year is before 2021. The lowest probability of 0.17 is observed when the repository size is less than 3 and the earliest publication year is 2021 or later. These rules suggest that repository size, the number of journals in the country of publication, and the amount of time a journal has been publishing significantly influence the likelihood that a JUOJS uses Crossref DOIs. These variables also exhibit a high relative importance (Table 5). Interestingly, variables such as language and discipline do not seem to play a significant role in predicting DOI usage.

Table 5. Relative importance of variables in relation to using DOIs

Variable	**Description**	**Relative importance**
Repository size	Number of journals in the repository	100.0
Earliest year publication	earliest year of first publication	98.6.
Open Page Rank of the journal	Page rank of the journal (1 - 10)	91.7
Number of journals in the country	number of journals in the focal journal's country of publication	83.4
Open Page Rank of the EndPoint	Page rank of the End Point (1 - 10)	70.8
Total number of documents	total number of documents since inception	54.9
GDP Per Capita	GDP per capita of the country of publication	13.2

Figure 8. Decision tree of rules to using Crossref DOIs*



*The tree model shows the rules for predicting coverage by OpenAlex and is based on the analysis of 47,625 journals. The numbers in each rectangle are the probability and percentage of journals (in brackets) meeting the criteria. The model achieved an accuracy of approximately 70%, which is an improvement of 42% (IMV = 0.42) over the prevalence baseline model. The precision score indicates that about 73% of the positive class predictions were accurate, suggesting a relatively low occurrence of false positives. In terms of recall, the model effectively identified 69% of actual positive cases, reflecting an ability to capture relevant instances. The F1 score of 71% suggests a good balance between precision and recall, indicating that the model performs well in both aspects. Additionally, the Area Under the Curve (AUC) of 0.75 highlights a moderate ability to differentiate between positive and negative classes across various threshold settings.

Discussion

Early visions of journal indexing were rooted in the ambitious goal of providing universal access to scientific knowledge and enabling researchers to discover, critique, and build upon prior work (Bradford 1934; Merton 1942). However, indexing databases such as Scopus and WoS have propelled exclusivity by curating lists of "quality" journals (Chavarro, Ràfols & Tang 2017), often marginalizing non-English, regional, or less "prestigious" outlets (van Leeuwen et al. 2001; Sivertsen & larsen 2012). Modern technologies and open infrastructures like OpenAlex have reinvigorated the goal of universal or inclusive indexing, but challenges remain in fully realizing this vision due to limitations in metadata quality, biases, and dependencies on established infrastructures (Alperin et al., 2024; Delgado-Quirós & Ortega, 2024; Culbert et al., 2024; Mongeon et al., 2023; Akbaritabar et al., 2023; Jahn et al., 2023; Alonso-Alvarez & van Eck, 2024). Our work supports efforts towards inclusive indexing by exploring how technical infrastructure requirements and established scholarly systems facilitate discoverability but also create systematic barriers to universal indexing.

Our results show that OpenAlex indexes 71% of JUOJS since 2020, which is up from 64% found for that year (Khanna et al. 2022). It was found that inclusion in OpenAlex is determined above all by a journal assigning a Crossref DOI for each article published. Doing so gives the journal a 97% probability of being indexed, with the final 3% likely missing due

to data mismatches. This is not surprising, given that, at present, being indexed in OpenAlex uses other indexes, including Crossref, for its sources. JUOJS are supported in their use of DOIs by a plugin that facilitates the registration and issuing of DOIs, along with instructional materials on this process (PKP 2024).

The pattern of Crossref DOI usage among JUOJS reflects elements of expense, experience, and culture involved in using this gateway metadata instrument. For instance, our findings show that 99% of JUOJS are produced in upper-middle (33,164 journals), high- (8,572 journals), and lower-middle-income countries (5,241 journals), while only 0.3% (165 journals) originate from low-income countries, suggesting that resource limitations in these regions hinder their ability to produce journals and adopt DOIs (see Figure 3). While noting the differences in the number of journals published by income group, on the whole, coverage levels are lower for the lower income countries: 70% of the JUOJS from high, upper-middle, and lower-middle income countries are indexed in OpenAlex, while only 47% of those from low-income countries are covered, in a pattern mapped out by other researchers (Beigel, 2024). The use of Crossref DOIs involves the expense of both an annual membership fee (in the area of $300 USD, for its lowest price tier) and then $1 per DOI issued (for new materials). Since 2014, PKP has used its standing as a Crossref Sponsoring Organization to enable JUOJS in low-income countries to join Crossref under PKP's membership, but this program has not had significant uptake. One outcome of this research will be to further motivate PKP's educational efforts in both understanding the barriers to, as well as encouraging, participation in mechanisms that lead to the indexing of content published in low-income countries.

The findings support editorial experience as an important factor in determining whether a journal is indexed and uses DOIs. This can be seen to be reflected in (a) the age of the journal (which also relates to its Open PageRank); (b) the number of journals per installation; and (c) the number of JUOJS published in the country. Each of these factors is seen to increase the likelihood that more knowledgeable editors and journal managers are involved in a given journal, leading to their decision to employ DOIs. This finding highlights the key role of an editorial culture in achieving universal indexing, but also a concentration of indexing on experienced journals in critical masses, hindering the visibility of works from journals in less resourceful contexts (Okune & Chan, 2023; Khurana et al., 2022; Chavarro, Ràfols, & Tang, 2018; van Bellen, Alperin, & Larivière, 2024).

Additionally, we found that JUOJS that publish in English are more likely to be found in OpenAlex (74% for monolingual and 71% for multilingual) compared to those without English (64% for monolingual and 55% for multilingual). This suggests that journals operating in English have a greater proximity to the means for indexing support, be it through great access to scholarly infrastructure or through better alignment with metadata standards. It also points to the concerted effort that will need to be made to extend that universality and communalism on a multilingual basis, as indexing systems have historically marginalized non-English languages and regional publications (Chavarro, Ràfols, & Tang, 2018; Cespedes et al., 2024). This pattern reflects broader systemic biases in indexing practices, where linguistic inclusivity remains a challenge despite the reinvigorated goal of universal indexing (Chavarro, 2017; Delgado-Quirós & Ortega, 2024).

A final issue to consider is the key role of Crossref as the principal path of indexing works from JUOJS in OpenAlex. This creates a centralization of power—the power to give visibility to content—on a system that, as we have shown, is not itself evenly distributed. With over 160M works, Crossref is a natural starting point for OpenAlex, as was their original reliance on the Microsoft Academic Graph database. Our analysis shows, however, the limitations of such an approach, and the continued need for all infrastructure providers and individuals publishers to make their metadata available in structured forms and for OpenAlex to continue to expand its indexing beyond MAG and Crossref (as it has done and continues to do). While not yet explored, the OpenAIRE plugin available for OJS may present an alternative for indexing that overcomes some of the barriers to the adoption of DOIs.

So while we would allow that there are reasons for the community to promote the advantages and means of using DOIs among journals, there are also reasons for all indexes to pursue ways of broadening their crawl to research beyond the literature involved in the dominant industry-led registration processes represented by Crossref. As shown by our tree in Figure 6, JUOJS not using Crossref DOIs still have a probability of being indexed by OpenAlex, which shows that OpenAlex's use of a diversity of data sources is effective in improving coverage. This might involve other DOI registration agencies, such as DataCite, which OpenAlex recently added, but clearly needs to go beyond a reliance on centralized databases and aggregators. Today, a large proportion of published research is discoverable online, at least in principle. Ascertaining whether an online document is a research publication, while not a trivial challenge, seems ready-made for machine learning models, such as Large Language Models, which could be trained to detect indexable materials on a multilingual basis. In addition, new standards for identifying research, such as PKP's Publication Facts Label for research articles, may also grow into an effective guide in web scraping for research indexing with potential use by all scholarly publishing platforms (Willinsky & Pimentel, 2023).

All told, we are encouraged by OpenAlex's interest in expanding its indexing strategies. This is further encouraged by a burgeoning collaboration between PKP and OpenAlex that will ensure full indexing of JUOJS in the near future, and seems like a promising means of pursuing the universal indexing ideals championed by Bradford. To enable researchers to more consistently consult *all* of the relevant and germane research in conducting their studies will, after all, only add to the integrity of the resulting claims. For this reason, we would conclude by observing that efforts at integrating all research into a well-indexed body of knowledge deserves a place amid the considerable attention being paid to strengthening research integrity at this point (Bolter 2024).

Limitations

Our work has required considerable effort to validate journal metadata, DOIs, classify journals into languages and disciplines, and merge different datasets. Given the multilingual nature of our dataset, language classifications and journal title validations against the ISSN.org have relied on different similarity measures that may miss some of the valid journals. However, we restricted our sample to make sure we are considering only journals whose metadata is consistent with a third-party authoritative source. In determining the main

language of a journal, we acknowledge that current multilingual algorithms to produce vectors work better in English, and as a result we may overestimate the presence of English language journals. In classifying the journals into disciplines, we have used the LLC subject classification, but we know that different classifications may lead to different results. When querying and merging datasets, different challenges appear such as the fast pace at which data sources change, potentially affecting our results. Our work only reflects the state of affairs to February 2024. We know, for example, that OpenAlex has since indexed works in DataCite, which will further index JUOJS that rely on this DOI registration agency.

Authors Contribution

D.C.: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing - original draft, and Writing - review & editing. **J.P.A.:** Conceptualization, Formal analysis, Funding acquisition, Methodology, Project administration, Resources, Supervision, Visualization, Writing - original draft, and Writing - review & editing. **J.W.:** Conceptualization, Funding acquisition, Resources, Supervision, Writing - original draft, and Writing - review & editing.

Acknowledgments

The authors would like to thank Saurabh Khanna, Jon Ball, Charles Rahal, Kyle Demes, and Mike Nason for their help with data collection and input into the research questions and comments to the preliminary versions of this document.

Competing Interests

The authors are associated with the Public Knowledge Project, developer of Open Journal Systems (DC is a Research Associate, JPA is the Scientific Director, and JW is the Founder).

Funding

This research was supported by the Social Sciences and Humanities Research Council (SSHRC) through grant #1007-2023-0001.

References

Alonso-Alvarez, P., & van Eck, N. J. (2024). Coverage and metadata availability of African publications in OpenAlex: A comparative analysis. *arXiv.* arXiv:2409.01120.

Alperin, J. P., Portenoy, J., Demes, K., Larivière, V., & Haustein, S. (2024). An analysis of the suitability of OpenAlex for bibliometric analyses. *arXiv*. https://doi.org/10.48550/arXiv.2404.17663

Ansorge, L. (2022). Hidden limitations of analyses via alternative bibliometric services. *Scientometrics, 128*(3), 2031-2033.

Beigel, F. (2024). Cartographies for an inclusive Open Science. *SciELO Preprints*. https://doi.org/10.1590/SciELOPreprints.10286

Bordignon, F. (2024). Is OpenAlex a revolution or a challenge for bibliometrics/bibliometricians? Groupe de travail Science Ouverte UDICE. https://enpc.hal.science/hal-04520837

Bouter, L. (2024). Why research integrity matters and how it can be improved. *Accountability in Research, 31*(8), 1277-1286.

Bradford, S. C. (1985). Sources of information on specific subjects. *Journal of Information Science, 10*(4), 173–180.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5â€“32. https://doi.org/10.1023/A:10109

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1983). *Classification and regression trees* (1st ed.). Routledge.

Cespedes, L., Kozlowski, D., Pradier, C., Sainte-Marie, M. H., Shokida, N. S., Benz, P., Poitras, C., Ninkov, A. B., Ebrahimy, S., Ayeni, P., Filali, S., Li, B., & Larivière, V. (2024). Evaluating the linguistic coverage of OpenAlex: An assessment of metadata accuracy and completeness. *arXiv*. https://arxiv.org/abs/2409.10633v2

Chavarro, D. (2017). *Universalism and particularism: Explaining the emergence and growth of regional journal indexing systems* [Doctoral dissertation, University of Sussex]. Sussex Research Repository. https://sussex.figshare.com/articles/thesis/Universalism_and_particularism_explaining_the_emergence_and_growth_of_regional_journal_indexing_systems/23440772/1/files/41151713.pdf

Chavarro, D., Ràfols, I., & Tang, P. (2018). To what extent is inclusion in the Web of Science an indicator of journal 'quality'? *Research Evaluation, 27*(2), 106-118.

Chavarro, D., & Alperin, J. (Forthcoming). Equity in Scholarly Visibility: Bridging the Gap for Journals using Open Journal Systems in OpenAlex. Berlin, STI 2024, 17-21 sept 2024

Chen, X., Wang, M., & Zhang, H. (2011). The use of classification trees for bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1*(1), 55-63.

Chtena, N., Alperin, J. P., Pinfield, S., Fleerackers, A., & Pasquetto, I (2024). Preprint servers and journals: Rivals or allies?. OSF. https://doi.org/10.31222/osf.io/fuydh

Choi, S. S., Cha, S. H., & Tappert, C. C. (2010). A survey of binary similarity and distance measures. *Journal of Systemics, Cybernetics and Informatics, 8*(1), 43-48.

COAR (2024). COAR launches task force on sustainable and interoperable PIDs models. https://coar-repositories.org/news-updates/coar-launches-task-force-on-sustainable-pids-models/

CommonCrawl Org. (2024). Common Crawl maintains a free, open repository of web crawl data that can be used by anyone. https://commoncrawl.org/

Culbert, J., Hobert, A., Jahn, N., Haupka, N., Schmidt, M., Donner, P., & Mayr, P. (2024). Reference coverage analysis of OpenAlex compared to Web of Science and Scopus. *arXiv*. https://doi.org/10.48550/arXiv.2401.16359

CWTS. (2024). CWTS Leiden Ranking Open Edition. https://open.leidenranking.com/

de Melo Maricato, J., Mazoni, A., Mugnaini, R., Packer, A. L., & Costas, R. (2023). SciELO as an open scientometric research infrastructure: General discussion of coverage in OpenAlex, WoS, Scopus and Dimensions. In *27th International Conference on Science, Technology and Innovation Indicators (STI 2023)*.

Delgado López-Cózar, E., Orduña-Malea, E., & Martín-Martí-n, A. (2019). Google Scholar as a data source for research assessment. In *Springer Handbook of Science and Technology Indicators* (pp. 95-127).

Delgado-Quirós, L., & Ortega, J. L. (2024). Completeness degree of publication metadata in eight free-access scholarly databases. *Quantitative Science Studies*, 1-19. https://doi.org/10.1162/qss_a_00286

DOAJ. (2024). List of journals. https://doaj.org/csv

DomCop. (2024). What is OpenPageRank? https://www.domcop.com/openpagerank/what-is-openpagerank

Domingue, B., Rahal, C., Faul, J., Freese, J., Kanopka, K., Rigos, A., ... & Tripathi, A. (2021). The InterModel Vigorish (IMV) as a flexible and portable approach for quantifying predictive accuracy with binary outcomes. https://doi.org/10.31235/osf.io/gu3ap

Edgar, B. D., & Willinsky, J. (2010). A survey of scholarly journals using Open Journal Systems. *Scholarly and Research Communication, 1*(2).

Facebook. (2024). English word vectors. https://fasttext.cc/docs/en/english-vectors.html

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861-874.

Hamming, R. W. (1950). Error detecting and error correcting codes. *Bell System Technical Journal, 29*(2), 147-160.

Hugar, J. G. (2019). Impact of open access journals in DOAJ: An analysis. *International journals of advanced library and information science, 7*(1), 448-455

Hückstädt, M. (2023). Ten reasons why research collaborations succeed.A random forest approach. *Scientometrics, 128*(3), 1923-1950.

Jaccard, P. (1912). The distribution of the flora in the alpine zone. *New Phytologist, 11*(2), 37-50.

Jiao, C., Li, K., & Fang, Z. (2023). How are exclusively data journals indexed in major scholarly databases? An examination of four databases. *Scientific Data, 10*(1), 737. https://doi.org/10.1038/s41597-023-02625-x

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of tricks for efficient text classification. *arXiv Preprint*. arXiv:1607.01759.

Kern, C., Klausch, T., & Kreuter, F. (2019). Tree-based machine learning methods for survey research. *Survey Research Method, 13*, 73â€“93. https://doi.org/10.18148/SRM/2019.V1I1.7395

Khanna, S., Ball, J., Alperin, J. P., & Willinsky, J. (2022). Recalibrating the scope of scholarly publishing: A modest step in a vast decolonization process. *Quantitative Science Studies, 3*(4), 912-930.

Khurana, P., Ganesan, G., Kumar, G., & Sharma, K. (2022). A bibliometric analysis to unveil the impact of digital object identifiers (DOI) on bibliometric indicators. In *Proceedings of Third International Conference on Computing, Communications, and Cyber-Security: IC4S 2021* (pp. 859-869). Springer Nature Singapore.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady, 10*(8), 707-710.

Lin, H., Lasser, J., Lewandowsky, S., Cole, R., Gully, A., Rand, D. G., & Pennycook, G. (2023). High level of correspondence across different news domain quality rating sets. *PNAS Nexus, 2*(9), pgad286.

Maddi, A., Maisonobe, M., & Boukacem-Zeghmouri, C. (2024). Geographical and Disciplinary Coverage of Open Access Journals: OpenAlex, Scopus and WoS. *arXiv preprint arXiv:2411.03325*.

Martín-Martí-n, A., Thelwall, M., Orduna-Malea, E., & Delgado López-Cózar, E. (2021). Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and

OpenCitations' COCI: A multidisciplinary comparison of coverage via citations. *Scientometrics, 126*(1), 871-906.

Merton, R. K. (1942). The ethos of science. *Journal of Legal and Political Sociology, 1*, 115-126.

Ohkura, N., Hirata, K., Kuboyama, T., & Harao, M. (2005). The q-gram distance for ordered unlabeled trees. In *Discovery Science: 8th International Conference, DS 2005, Singapore, October 8â€“11, 2005. Proceedings 8* (pp. 189-202). Springer Berlin Heidelberg.

Okune, A., & Chan, L. (2023). Digital Object Identifier: Privatising knowledge governance through infrastructuring. In *Routledge Handbook of Academic Knowledge Circulation* (pp. 278-287). Routledge.

Our Research. (2024). Sorbonne University announces switch to OpenAlex. https://blog.ourresearch.org/sorbonne-university-announces-switch-to-openalex/

Owen, B., & Stranack, K. (2012). The Public Knowledge Project and Open Journal Systems: Open source options for small publishers. *Learned Publishing, 25*(2), 138-144.

PKP (2024). Getting started with DOIs. https://docs.pkp.sfu.ca/crossref-ojs-manual/en/gettingStarted

Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv Preprint*. arXiv:2205.01833.

Reimers, N. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-Networks. *arXiv Preprint*. arXiv:1908.10084.

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval*. McGraw-Hill.

Sivertsen, G., & Larsen, B. (2012). Comprehensive bibliographic coverage of the social sciences and humanities in a citation index: An empirical analysis of the potential. *Scientometrics, 91*(2), 567â€“575. http://doi.org/10.1007/s11192-011-0615-3

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, *28*(1), 11-21.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics, 9*(1), 307. https://doi.org/10.1186/1471-2105-9-307

van Bellen, S., Alperin, J. P., & Larivière, V. (2024). The oligopoly of academic publishers persists in exclusive database. *arXiv Preprint*. arXiv:2406.17893.

van Leeuwen, T., Moed, H., Tijssen, R., Visser, M., & van Raan, A. (2001). Language biases in the coverage of the Science Citation Index and its consequences for international comparisons of national research performance. *Scientometrics, 51*(1), 335â€“346.

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics, 7*, 1-8.

Wang, K., Shen, Z., Huang, C., Wu, C.-H., Dong, Y., & Kanakia, A. (2020). Microsoft academic graph: When experts are not enough. *Quantitative Science Studies, 1*(1), 396â€“413. https://doi.org/10.1162/qss_a_00021

Wang, Z. (2022). Use of supervised machine learning to detect abuse of COVID-19 related domain names. *Computers and Electrical Engineering, 100*, 107864.

Winkler, W. E. (1990). String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods, American Statistical Association* (pp. 354-359).

Willinsky, J., & Pimentel, D. (2024). The publication facts label: A public and professional guide for research articles. *Learned Publishing*. https://doi.org/10.1002/leap.1599

World Bank. (2024a). Classifying countries by income. https://datatopics.worldbank.org/world-development-indicators/stories/the-classification-of-countries-by-income.html

World Bank. (2024b). World Development Indicators. https://databankfiles.worldbank.org/public/ddpext/?f=1&v=1

Zhang, L., Cao, Z., Shang, Y., Sivertsen, G., & Huang, Y. (2024). Missing institutions in OpenAlex: Possible reasons, implications, and solutions. *Scientometrics*. https://doi.org/10.1007/s11192-023-04923-y

Appendix 1

Table A1. Descriptive statistics for the variables used in Classification tree 1.

OpenAlex	Not in OpenAlex			In OpenAlex		
Variable	**N**	**Mean**	**SD**	**N**	**Mean**	**SD**
countryNumJournals	**13971**	**9772**	**10217**	**33654**	**11134**	**10363**
Scopus	13971			33654		
... 0	13677	98%		30266	90%	
... 1	294	2%		3388	10%	
DOAJ	13971			33654		
... 0	13116	94%		25902	77%	
... 1	855	6%		7752	23%	
CrossRef	13971			33654		
... 0	13045	93%		11953	36%	
... 1	926	7%		21701	64%	
DataCite	13971			33654		
... 0	13379	96%		33460	99%	
... 1	592	4%		194	1%	
Medra	13971			33654		
... 0	13862	99%		33618	100%	
... 1	109	1%		36	0%	
JALC	13971			33654		
... 0	13970	100%		33654	100%	
... 1	1	0%		0	0%	
Airiti	1397			33654		

	1					
... 0	13970	100%		33654	100%	
... 1	1	0%		0	0%	
openPageRankDecEPAvg	13971	2.7	1.4	33654	3.2	1.2
openPageRankDecAvg	13971	1.4	1.7	33654	2	1.9
RepoSize	13971	58	178	33654	37	92
earliestYearPub	13971	2019	4.2	33654	2018	11
numDocsTotal	13971	169	557	33654	253	586
GDPPerCapita	13971	11609	16176	33654	12131	17196
lang	13971			33654		
... NA	123	1%		66	0%	
... Mono-Eng	3033	22%		9024	27%	
... Mono-NonEng	2914	21%		5190	15%	
... Multi-Eng	7462	53%		18819	56%	
... Multi-NonEng	439	3%		555	2%	
MainSubject	13971			33654		
... Agriculture	319	2%		1060	3%	
... Auxiliary sciences of history	38	0%		66	0%	
... Bibliography. Library science. Information resources	32	0%		61	0%	
... Education	1646	12%		3444	10%	
... Fine Arts	212	2%		451	1%	
... General Works	8	0%		20	0%	
... Geography. Anthropology. Recreation	327	2%		929	3%	

... History (General) and history of Europe	147	1%		390	1%	
... History America	3	0%		11	0%	
... Language and Literature	845	6%		2116	6%	
... Law	472	3%		1161	3%	
... Medicine	2384	17%		6081	18%	
... Military Science	5	0%		14	0%	
... Music and books on Music	26	0%		56	0%	
... Naval Science	0	0%		1	0%	
... Philosophy. Psychology. Religion	1054	8%		2402	7%	
... Political science	119	1%		318	1%	
... Science	1605	11%		4374	13%	
... Social Sciences	3102	22%		6730	20%	
... Technology	1627	12%		3969	12%	

Appendix 2.

Table A2. Descriptive statistics for the variables used in classification tree 2.

CrossRef	Not in CrossRef			In CrossRef		
Variable	**N**	**Mean**	**SD**	**N**	**Mean**	**SD**
countryNumJournals	24998	8425	9932	22627	13287	10177
openPageRankDecEPAvg	24998	2.8	1.4	22627	3.4	1.1
openPageRankDecAvg	24998	1.4	1.8	22627	2.2	1.9
RepoSize	24998	51	152	22627	35	82
earliestYearPub	24998	2019	12	22627	2017	4.8
numDocsTotal	24998	223	611	22627	234	541
GDPPerCapita	24998	13180	17925	22627	10650	15593
lang	24998			22627		
... lang_not_available	189	1%		0	0%	
... Monolingual_English	6955	28%		5102	23%	
... Monolingual_non_English	4765	19%		3339	15%	
... Multilingual_English	12418	50%		13863	61%	
... Multilingual_non_English	671	3%		323	1%	
MainSubject	24998			22627		
... Agriculture	645	3%		734	3%	
... Auxiliary sciences of history	65	0%		39	0%	
... Bibliography. Library science. Information	59	0%		34	0%	

resources						
... Education	2585	10%		2505	11%	
... Fine Arts	349	1%		314	1%	
... General Works	22	0%		6	0%	
... Geography. Anthropology. Recreation	653	3%		603	3%	
... History (General) and history of Europe	297	1%		240	1%	
... History America	10	0%		4	0%	
... Language and Literature	1476	6%		1485	7%	
... Law	781	3%		852	4%	
... Medicine	4797	19%		3668	16%	
... Military Science	11	0%		8	0%	
... Music and books on Music	54	0%		28	0%	
... Naval Science	0	0%		1	0%	
... Philosophy. Psychology. Religion	1663	7%		1793	8%	
... Political science	232	1%		205	1%	
... Science	3157	13%		2822	12%	
... Social Sciences	5243	21%		4589	20%	
... Technology	2899	12%		2697	12%	

This preprint was submitted under the following conditions:

- The authors declare that they are aware that they are solely responsible for the content of the preprint and that the deposit in SciELO Preprints does not mean any commitment on the part of SciELO, except its preservation and dissemination.
- The authors declare that the necessary Terms of Free and Informed Consent of participants or patients in the research were obtained and are described in the manuscript, when applicable.
- The authors declare that the preparation of the manuscript followed the ethical norms of scientific communication.
- The authors declare that the data, applications, and other content underlying the manuscript are referenced.
- The deposited manuscript is in PDF format.
- The authors declare that the research that originated the manuscript followed good ethical practices and that the necessary approvals from research ethics committees, when applicable, are described in the manuscript.
- The authors declare that once a manuscript is posted on the SciELO Preprints server, it can only be taken down on request to the SciELO Preprints server Editorial Secretariat, who will post a retraction notice in its place.
- The authors agree that the approved manuscript will be made available under a Creative Commons CC-BY license.
- The submitting author declares that the contributions of all authors and conflict of interest statement are included explicitly and in specific sections of the manuscript.
- The authors declare that the manuscript was not deposited and/or previously made available on another preprint server or published by a journal.
- If the manuscript is being reviewed or being prepared for publishing but not yet published by a journal, the authors declare that they have received authorization from the journal to make this deposit.
- The submitting author declares that all authors of the manuscript agree with the submission to SciELO Preprints.