Data Dictionary

This file describes the fields of the dataset of Journals Using OJS (JUOJS) analyzed in the paper "On the Open Road to Universal Indexing: OpenAlex and Open Journal Systems", by Chavarro, Alperin, and Willinsky (2024).

1. ISSN_OJS

• The primary International Standard Serial Number extracted from the OJS Beacon metadata. This is the ISSN provided by the journal's own OJS record, validated against the official ISSN registry.

2. ISSN_OJS2

• A secondary ISSN identifier (e.g. the e-ISSN or an additional ISSN variant) as recorded in the OJS metadata. The ".x" suffix denotes that this field is part of the versioned supplementary dataset.

3. journal_name

• The title of the journal as registered in the OJS system. This name is taken directly from the publisher-provided metadata—and it may have been standardized (e.g., via transliteration and phonetic matching) during validation.

4. country_consolidated

• The final or "consolidated" country of publication for the journal. This value is determined by aggregating several sources (such as ISSN metadata, DOAJ listings, and IP-address based geolocation) to assign a country affiliation.

5. record_count_2020

• The number of "records" (i.e. items or metadata entries) for the journal from the year 2020 as captured from the OJS Beacon. (A journal is considered "active" for a year if it published at least five items and was validated against ISSN.org.)

6. record_count_2021

• The analogous count of records for the year 2021.

7. record_count_2022

• The analogous count of records for the year 2022.

8. record_count_2023

• The analogous count of records for the year 2023.

9. numDocs2020

• The number of scholarly documents (e.g. articles, papers) published by the journal in 2020. This count is used to assess annual publication volume.

10. numDocs2021

• The number of documents published in 2021.

11. numDocs2022

• The number of documents published in 2022.

12. numDocs2023

• The number of documents published in 2023.

13. OpenAlex

• A field indicating whether the journal is indexed in OpenAlex. (a binary indicator, where "1" means that at least one validated ISSN appears in OpenAlex, and "0" means it is not covered.)

14. incomeGroup

• The World Bank–derived income group of the journal's country (e.g. High, Upper-Middle, Lower-Middle, or Low income). This classification is used to study systematic differences in indexing and DOI usage across country income levels.

15. GDPPerCapita

• The Gross Domestic Product per capita (sourced from World Bank data) corresponding to the journal's country. This numeric indicator is used as a proxy for resource availability and economic context.

16. countryNumJournals

• The total number of JUOJS (journals using OJS) identified in the same country. This field captures the "critical mass" of journals per national setting.

17. indScopus

• A binary indicator denoting whether the journal is indexed in Scopus (1 = yes; 0 = no).

18. indDOAJ

• A binary indicator showing whether the journal is included in the Directory of Open Access Journals (DOAJ).

19. indDOI

• A binary indicator that flags if the journal assigns Digital Object Identifiers (DOIs) to its publications (typically indicating whether a Crossref DOI is present).

20. CrossRefMatchNumDocsFinal

• The final count of the journal's documents that were successfully matched against Crossref DOI records. This value reflects how many items have verifiable Crossref-registered DOIs.

21. DataCiteMatchNumDocs

• The number of documents from that journal matched with DOI records from DataCite, representing an alternative DOI registration agency.

22. mEDRAMatchNumDocs

• The number of journal documents that were matched to records in Medra (another DOI registration or metadata validation system, if used). If Medra is not applicable to all cases, a zero or null value would indicate no matching.

23. JaLCMatchNumDocs

• The count of documents matched to records from the JALC system. (JALC is used to capture DOIs from certain regional or subject-specific registries and provides an alternative pathway for indexing.)

24. AiritiMatchNumDocs

• The number of documents for which a match was found using the Airiti system, another regional DOI/metadata registry.

25. CrossRef

• A field indicating the presence of DOIs registered with Crossref. (expressed as a binary indicator that the journal uses Crossref's service for its DOIs, given that nearly all validated DOIs in the sample come from Crossref.)

26. DataCite

• A field indicating the presence of DOIs registered via DataCite for the journal.

27. Medra

• A marker showing whether the journal has any DOIs registered through Medra. This is expressed as a binary indicator.

28. JALC

• A binary indicator of whether the journal uses the JALC registration mechanism for DOIs.

29. Airiti

• A binary indicator of whether the journal has DOIs registered with the Airiti service.

30. DOIs

• The overall total count of Digital Object Identifiers assigned to the journal's published documents, as collected from the journal's metadata.

31. matchedDOIs

• The number of DOIs that successfully "matched" (i.e. were validated against external DOI resolution services such as DOI.org and Crossref), indicating the completeness and correctness of the journal's DOI metadata.

32. RepoSize

• The size of the repository (i.e. the OJS installation) that hosts the journal. It is defined by the number of journals managed on that same installation and is used as an indicator of editorial infrastructure scale.

33. earliestYearPub

• The year in which the journal published its first document. This field provides a measure of the journal's age and publishing history.

34. numDocsTotal

• The cumulative number of documents published by the journal since its inception. This aggregation spans all available years in the dataset.

35. openPageRankDecAvg

• The average Open PageRank score (ranging from 0 to 10) of the journal's webpage. This score, derived from the CommonCrawl dataset, indicates the digital visibility or "web authority" of the journal.

36. openPageRankDecEPAvg

• The average Open PageRank score for the OJS installation's endpoint (the main site or repository home page). This metric reflects the web reputation of the entire journal portal, not just a single journal's page.

37. lang

• The language profile of the journal's published articles, as determined by automated language detection (using Meta's FastText model). Journals are categorized into groups (e.g., monolingual English, monolingual non-English, multilingual with English, multilingual without English, or not available).

38. percentage_english

• The percentage of the journal's articles that are published in English. This quantitative measure is derived from aggregating individual article language classifications.

39. percentage_dominant_non_english

• The percentage of articles for which a non-English language is dominant (for instance, when more than 90% of a journal's output is in a single non-English language). This field helps distinguish journals that are primarily non-English.

40. MainSubject

• The primary scholarly discipline or subject area of the journal. This classification is produced by applying a machine-learning model (trained on DOAJ subject categories and multilingual embeddings of journal titles) and corresponds to top-level categories (e.g., Medicine, Science, Social Sciences, etc.).

These definitions should provide clarity for dataset users on what each variable represents, how it was derived, and its role in the analysis of journal indexing, DOI usage, and global scholarly visibility as described in the paper.