

README for Dataset Files

01_merged_data.xlsx
02_merged_data_final.xlsx

Version 1.0

Overview

This README covers two related datasets derived from a merge of bibliometric data from Scopus and Web of Science (WoS).

The first dataset, `01_merged_data.xlsx`, includes records with duplicates removed. The second dataset, `02_merged_data_final.xlsx`, is a refined version of the first, where records were further manually screened to ensure relevance and accuracy.

Both datasets are presented in Excel file formats and retain key bibliometric details for analysis.

File Formats

`01_merged_data.xlsx`

- **File Type:** Excel (.xlsx)
- **Total Records:** 759
- **Total Columns (Fields):** 33

`02_merged_data_final.xlsx`

- **File Type:** Excel (.xlsx)
- **Total Records:** 608
- **Total Columns (Fields):** 33

Description of Key Columns

This dataset includes a comprehensive set of metadata fields, with the following key columns:

- **AU:** Authors - Names of the authors.
- **DE:** Authors' Keywords - Keywords provided by the authors.
- **ID:** Keywords Plus[®] - Additional keywords derived from the titles of cited references.
- **C1:** Author Address - Addresses of the authors.
- **JI:** ISO Source Abbreviation - The ISO abbreviation for the source title.
- **AB:** Abstract - A summary of the document's content.
- **RP:** Reprint Address - Address for requesting reprints of the publication.
- **DI:** Digital Object Identifier (DOI) - A unique identifier for the document.
- **BE:** Editors - Editors of the publication.
- **BN:** International Standard Book Number (ISBN) - The ISBN for books.
- **SN:** International Standard Serial Number (ISSN) - The ISSN for the journal.
- **SO:** Publication Name - The name of the journal or publication.
- **LA:** Language - The language in which the document is written.
- **TC:** WoS Core Collection Times Cited Count - The number of times the document has been cited in the WoS Core Collection.
- **PN:** Part Number - The part number of the document.
- **PU:** Publisher - The publisher of the document.
- **DB:** Database Source - The source of the bibliometric data (e.g., WoS, Scopus).
- **TI:** Document Title - The title of the document.
- **DT:** Document Type - Type of document (e.g., article, review).
- **url:** URL - The URL for accessing the record or document.
- **VL:** Volume - The volume number of the publication.
- **PY:** Year Published - The year of publication.
- **J9:** 29-Character Source Abbreviation - Abbreviation of the source title.
- **CR:** Cited References - References cited by the document.

Data Cleaning Process

- **Manual Screening:** The second dataset (`02_merged_data_final.xlsx`) underwent additional manual screening to remove irrelevant or inaccurate records and ensure data quality.
- **Metadata Verification and Correction:** Corrections were made to key metadata fields, including:
 - **Authors' Keywords (DE):** Keywords were reviewed and corrected to ensure accuracy.
 - **Cited References (CR):** References were checked and updated for completeness and correctness.
 - Other fields were also checked and corrected as necessary to ensure consistency and reliability.

Limitations

- **Data Incompleteness:** Some records may have incomplete metadata fields due to inconsistencies in the original data sources.
- **Potential Bias:** The datasets may reflect biases inherent in the original data sources (Scopus and WoS), such as coverage bias or language bias.
- **Manual Screening Subjectivity:** The manual screening process may introduce subjective judgment in the inclusion or exclusion of records.

Date of Data Extraction

The source files (`wos_data.bib` and `scopus_data.csv`) were extracted from the Scopus and WoS databases on **31 August 2023**.

Usage

These datasets are suitable for detailed bibliometric analysis, including research trends, authorship patterns, and citation studies. They provide a comprehensive and curated collection of scholarly records, useful for researchers, librarians, and analysts interested in scholarly communication and research impact.

Acknowledgments

The datasets were compiled using the `bibliometrix` package in R and refined through a combination of automated and manual review processes. Special thanks to the creators of the `bibliometrix` and `openxlsx` packages, as well as to the individuals involved in the manual screening process.