The Journal of Open Source Software

Bibliography Verification Tool: Automated Reference Validation Using CrossRef and PubMed

P. V. Sundar Balakrishnan[1]

1 University of Washington Bothell, USA

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Open Journals

Reviewers:

- @openjournals

Submitted: 01 January 1970
Published: unpublished

Summary

Accurate bibliographic references are essential for scientific integrity, yet verifying them manually—particularly in manuscripts with large reference lists—is laborious and error-prone. The Bibliography Verification Tool provides automated, reproducible validation of references against authoritative databases. This Python-based system extracts citations from Microsoft Word documents, queries CrossRef and PubMed APIs, evaluates metadata consistency using fuzzy matching, and generates detailed verification reports with confidence scoring. The tool accommodates diverse reference types, including journal articles, books, classic editions, and ancient texts, and integrates seamlessly with R for quantitative assessment. It offers researchers, reviewers, and editors an efficient and transparent method for ensuring reference accuracy in academic publishing. The software is archived at Zenodo and assigned a permanent DOI (Balakrishnan, 2025).

Statement of Need

Reference accuracy affects literature discoverability, citation tracking, and the credibility of scholarly work. Metadata errors—such as incorrect publication years, misattributed authorship, incorrect or malformed DOIs, or incomplete metadata—remain common and often unnoticed until the peer-review process. These mistakes consume reviewer time and can compromise the perceived rigor of a manuscript (Moed, 2005).

Popular citation managers (e.g., Zotero, EndNote, Mendeley) are highly effective for organizing references but do not validate metadata against external databases (Kratochvil, 2011). Researchers typically rely on ad hoc manual checks using CrossRef or PubMed, a process that becomes infeasible when bibliographies contain 50–300 references, as is common in review articles, dissertations, and meta-analyses. While Python libraries such as `habanero`, `crossrefapi`, and `biopython` offer programmatic access to bibliographic APIs, using these tools requires custom scripting and does not provide an end-to-end workflow.

The Bibliography Verification Tool fills this methodological gap with a turnkey solution for automated metadata validation. It supports three primary use cases: (1) pre-submission manuscript preparation, enabling authors to verify and correct references before journal submission; (2) peer-review and editorial quality control, providing reviewers and editors a consistent way to evaluate bibliography integrity; and (3) reproducibility audits, where researchers examine citation accuracy across multiple publications. By automating extraction, matching, and reporting, the tool reduces human error and ensures transparent, reproducible verification.

Description of the Software

The tool implements a four-stage verification pipeline: extraction, query, matching, and reporting.

Extraction

The system reads Microsoft Word (.docx) files using `python-docx` and extracts APA-formatted citations using a sequence of regular expressions. It identifies reference types—journal articles, books, classic works with original publication years, and ancient texts—and extracts authors, publication years, titles, and DOIs. Unicode normalization ensures correct handling of diacritics (e.g., Treviño → Trevino). Extraction failures (e.g., missing titles or ambiguous patterns) are logged to assist users in adjusting problematic references.

Classic editions are handled explicitly. The tool detects expressions such as "Original work published 1785" and records both the edition year and the original publication date. Works published before 1800 are automatically classified as ancient texts, which are excluded from automated database queries because modern metadata sources do not index them.

Query

Extracted references are validated against the CrossRef REST API (Hendricks et al., 2020), with title- and author-based searches augmented by publication year filters (±2 years to account for differences between online and print publication dates). For biomedical references, the tool also consults PubMed via E-utilities. All queries follow API etiquette recommendations, including the use of contact email headers, polite rate limiting (approximately one request per second), and exponential backoff. A persistent session with retry logic ensures robust operation even during network fluctuations.

Matching and Scoring

A composite match score (0–100) evaluates the consistency between extracted metadata and database results across three dimensions:

- **Title similarity (50 points)**: measured using fuzzy string matching with type-specific thresholds (0.85 for journal articles, 0.75 for books to accommodate subtitles).
- **Year alignment (25 points)**: full credit for exact matches or tolerance within ±2 years.
- **Author match (25 points)**: based on normalized comparison of first-author surnames.

References with scores 50 are labeled VERIFIED. Lower scores trigger NEEDS_REVIEW status, accompanied by detailed issue flags (e.g., YEAR_MISMATCH, LOW_MATCH_CONFIDENCE, NO_DOI_FOUND). Classic editions receive specialized handling: original publication years are reported but not treated as mismatches for modern editions.

Output and Analysis

The tool generates three complementary outputs:

1. **verification_report.csv**: detailed metadata and match scores for archival and review.
2. **verification_log.txt**: human-readable summary prioritizing items requiring attention.
3. **verification_for_R.csv**: R-ready file including boolean filters such as `Needs_Manual_Check`, `High_Confidence`, and `Is_Book`.

A companion R script (`analyze_verification_results.R`) provides 10 pre-built analysis functions for calculating verification rates, exploring issue categories, and producing publication-ready visualizations via `tidyverse` and `ggplot2`. This integration facilitates reproducible reporting and aids meta-researchers studying bibliographic quality.

State of the Field

Reference verification tools fall into three categories:

Citation Management Software. Tools such as Zotero (Kratochvil, 2011), EndNote, and Mendeley excel at organizing and formatting references but do not validate metadata. They often import metadata from publishers without checking its correctness.

API Libraries. Python packages including `habanero`, `crossrefapi`, and `biopython` provide programmatic access to CrossRef and PubMed (Cock et al., 2009), but require significant programming skill to integrate extraction, matching, and reporting. They are designed for developers, not for researchers seeking an end-user workflow.

Manual Verification. Researchers often check problematic references manually using CrossRef or PubMed. This approach is time-intensive, inconsistent, and does not generate reproducible records of verification decisions.

The Bibliography Verification Tool bridges these gaps by providing a cohesive, user-friendly pipeline that automates extraction, fuzzy matching, issue flagging, and reporting. Its treatment of classic editions and ancient texts addresses bibliographic edge cases frequently encountered in the humanities and social sciences. The integration with R further supports transparency initiatives in reproducibility research (Hardwicke & Ioannidis, 2018), enabling systematic evaluation of bibliography quality.

Acknowledgements

Development of the tool benefited from experience verifying large bibliographies in review articles and meta-analyses. I am grateful for the public APIs provided by CrossRef and PubMed, and for the R community's `tidyverse` and `ggplot2` packages that support downstream analysis.

References

[The bibliography is maintained separately in **paper.bib** as required by JOSS.]

Balakrishnan, P. V. (Sundar). (2025). *Bibliography verification tool: Automated reference verification against CrossRef and PubMed* (Version 1.0.1). Zenodo. https://doi.org/10.5281/zenodo.17622390

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, B., Wilczynski, B., & Hoon, M. J. L. de. (2009). Biopython: Freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163

Hardwicke, T. E., & Ioannidis, J. P. A. (2018). Populating the data ark: An attempt to retrieve, preserve, and liberate data from the most highly-cited psychology and psychiatry articles. *PLOS ONE*, *13*(8), e0201856. https://doi.org/10.1371/journal.pone.0201856

Hendricks, G., Bragazzi, N. L., Santamaria, P., & Planesi, S. (2020). The crossref REST API. *Data Science Journal*, *19*, 19. https://doi.org/10.5334/dsj-2020-019

Kratochvil, D. (2011). Zotero: A next-generation reference manager. *The Charleston Advisor*, *13*(1), 32–34. https://doi.org/10.5260/chara.13.1.32

Moed, H. F. (2005). *Citation analysis in research evaluation*. Springer. https://doi.org/10.1007/1-4020-3714-5