

SciFilter: A Metadata-Driven Approach to Registry-Aligned Screening of Scholarly Publication Sources at Scale

Abiodun Afolayan Ogunyemi

`abiodun.a.ogunyemi@gmail.com`

Tallinn University

Merja Bauters

Tallinn University

Emmanuel Ayo Orisadare

University of Ilesha

Research Article

Keywords: Bibliometrics. Quality assessment. Publication screening. Inter-rater agreement. JUFO index. Norwegian register

Posted Date: October 21st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11120204/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

SciFilter: A Metadata-Driven Approach to Registry-Aligned Screening of Scholarly Publication Sources at Scale

Abiodun Afolayan Ogunyemi^{1*}, Merja Bauters¹, and Emmanuel Ayo Orisadare²

¹School of Digital Technologies, Tallinn University, Estonia

²Department of Information and Communication Technology, University of Ilesa, Nigeria

*Corresponding author: abiodun.a.ogunyemi@gmail.com

SciFilter: A Metadata-Driven Approach to Registry-Aligned Screening of Scholarly Publication Sources at Scale

Abstract

Quality assessment is vital to research evaluation, yet accessible tools for structured, automated screening of publication sources remain limited. Existing services are either proprietary or lack automated batch-processing capabilities. This study introduces SciFilter, an open-access system designed to automate publication-source screening by aligning bibliographic metadata with established national frameworks—specifically the Finnish Publication Forum (JUFO) and the Norwegian Register for Scientific Journals, Series and Publishers (Norwegian Register). We evaluate the system using a dual-pathway design: first, establishing a human-curated benchmark dataset (N=500) across indexed and non-indexed sources, with reliability confirmed via Cohen's Kappa ($\kappa=0.95$, between human raters) and Fleiss' Kappa ($\kappa=0.94$, across both human raters and SciFilter); and subsequently measuring system performance against the harmonised gold standard (precision=0.95, recall=0.96, accuracy=0.96). Confusion matrix analyses show that most classification errors occurred at adjacent registry-level boundaries, consistent with boundary ambiguity; a smaller share involved larger tier discrepancies (e.g., Level 3 misclassified as Level 0), which merit closer review. By ingesting bulk bibliographic inputs (Excel, CSV, XML), SciFilter provides a scalable, reproducible first-pass filtering mechanism and accessible guardrail for systematic literature reviews, institutional portfolio audits, and research assessment workflows where manual registry lookups are unfeasible.

Keywords: Bibliometrics . Quality assessment . Publication screening . Inter-rater agreement . JUFO index . Norwegian register

Introduction

Systematic literature reviews and bibliometric analyses require rigorous screening to ensure the integrity and reproducibility of the publication venues included (Kitchenham & Charters, 2007). High-quality academic outputs are essential building blocks for both future scholarship and industrial innovation, as evidenced by the reliance of corporate research and development on public, peer-reviewed literature (Hicks et al., 2000; Hogg & Smith, 1959; Lopes-Bento & Simeth, 2024; Narin et al., 1997). However, the rapid expansion and growing heterogeneity of scholarly publishing have made it difficult for researchers and institutional administrators to efficiently assess the credibility and suitability of publication sources.

This challenge is exacerbated by the growth of questionable or low-transparency publication venues—often discussed under the contested term "predatory journals" (see, for example, do Carmo et al., 2026)—making careful venue selection critical to maintaining the integrity of archival knowledge. While established proprietary databases such as Web of Science and Scopus index reputable venues, sources are periodically reviewed, demoted, or delisted, affecting their visibility and perceived legitimacy. At the same time, subscription paywalls limit access and exercise indexing controls, whereas inclusive platforms like Google Scholar capture broader literature without strict editorial quality gates, leaving the burden of source verification entirely to the researcher.

To support researchers and institutional evaluators, expert-curated national registry systems have emerged as structured, field-sensitive frameworks for assessing publication channels (Hojnik, 2021). Systems such as the JUFO and the Norwegian Register assess channels through discipline-specific expert panels (Pölönen, 2019; Pölönen & Pylvänäinen, 2023). Publication sources are classified on a tiered scale (JUFO Levels 0–3; Norwegian Levels 0–2), with higher tiers reflecting peer-review rigour, editorial diversity, and domain recognition

rather than raw citation volume alone. A Level 0 rating, or absence from these systems, indicates that a source falls outside expert-reviewed coverage or national inclusion criteria, not an explicit evaluation of poor quality.

Despite their value in operationalising Responsible Research Assessment (RRA) principles (Hicks et al., 2015; San Francisco Declaration on Research Assessment (DORA, 2013)), national registries primarily function as web portals for manual, single-record lookups. This manual lookup interface creates a severe operational bottleneck for tasks requiring bulk source screening, such as institutional research portfolio audits, systematic review pipelines, or large-scale bibliometric analyses.

To bridge this operational gap, we developed SciFilter, an automated, metadata-driven screening tool. SciFilter ingests raw bibliographic records (Excel, CSV, XML), normalises source metadata, and maps them directly against releases of the centralised reference databases for the JUFO and the Norwegian registries. Beyond bulk file uploads, SciFilter provides actionable decision support for early-career researchers and students navigating source evaluation, serving as an accessible guardrail for responsible research assessment without requiring deep manual bibliometric expertise.

This study evaluates the performance and operational limits of SciFilter through three primary research questions:

RQ1: To what extent can an automated, metadata-driven lookup reproduce expert-curated registry classifications across diverse publication channels?

RQ2: What types of classification discrepancies arise in registry-aligned screening, and what structural patterns do they reveal about the boundary conditions of registry-based evaluation systems?

RQ3: How can metadata-driven registry alignment enable scalable, reproducible screening within bibliometric and research evaluation workflows?

To provide a structured overview of the approach, Figure 1 presents a conceptual model of metadata-driven registry alignment. The model formalises the process by which bibliographic metadata are transformed into registry-aligned classifications. At the input layer, heterogeneous bibliographic metadata (e.g., journal names, ISSNs) are ingested and standardised. These standardised records are then mapped to registry-based classification systems, such as JUFO and the Norwegian Register, which serve as the evaluative reference layer. The alignment process produces classification outputs that indicate indexing status and registry levels.

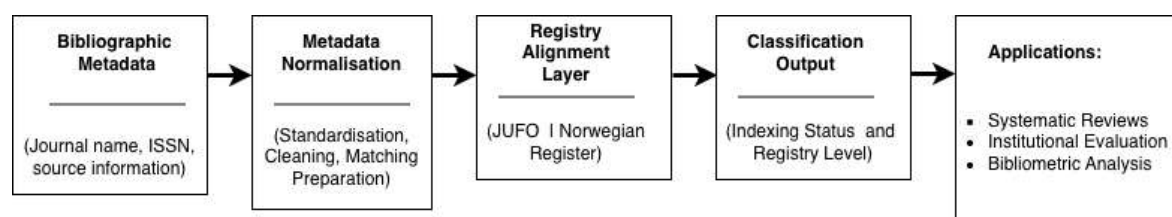


Fig. 1 Conceptual model of metadata-driven registry alignment for screening scholarly publication sources.

The model is not specific to SciFilter and can be extended to other registry-based or indexing systems. It also highlights the relationship between classification outputs and their application contexts, including systematic literature reviews, institutional research evaluation, and bibliometric analysis. In this sense, the model does not define a new evaluation metric but rather describes how existing expert-based classification systems can be implemented through structured metadata processes. This abstraction enables the approach to be generalised beyond the specific implementation presented in this study.

The rest of the article is organised as follows: After the introduction, the article is divided into six key sections. The ‘Quality Assessment of Scholarly Publications’ section offers a critical review of established and emerging metrics for evaluating research output, emphasising their advantages, criticisms, and relevance to responsible research assessment. The ‘SciFilter System Architecture’ section describes the design and functional components of the SciFilter tool, outlining its technical infrastructure and quality-screening workflow. The Methods section explains the evaluation framework, data sources, and metrics used to assess SciFilter’s reliability and accuracy. The Results section presents the evaluation findings, including confusion matrix analyses and key performance indicators. In the Discussion section, the implications of the results are interpreted in the context of the existing literature, the study’s limitations are candidly discussed, and future research on assessing publication quality is outlined. Finally, the Conclusion section summarises the study’s contributions and suggests directions for system improvements.

Quality Assessment of Scholarly Publications

Evaluating the academic standing and integrity of scholarly publication channels remains one of the most contested domain challenges within bibliometrics and research governance (Cronin & Sugimoto, 2014; Hicks et al., 2015). Historically, institutional evaluation relied on aggregate quantitative citation metrics, most prominently the Journal Impact Factor (JIF). However, extensive critical scholarship has demonstrated that citation-based metrics suffer from severe structural biases: they are highly skewed by extreme outlier papers, heavily favour English-language STEM disciplines over the Social Sciences and Humanities (SSH), and are vulnerable to strategic gaming, such as coercive self-citation and editorial cartels (Bollen et al., 2009; Seglen, 1997).

These concerns were not new: bibliometric scholarship had already called for citation analysis to be applied transparently, within a clear policy context, and as a supplement to expert judgment rather than a substitute for it (Moed, 2005). To counter the misapplication of uncontextualised metrics, the scientific community articulated normative guidelines through the San Francisco Declaration on Research Assessment (DORA, 2013) and the Leiden Manifesto (Hicks et al., 2015). These frameworks establish that quantitative indicators must support, rather than supplant, expert qualitative assessment and account for field-specific publishing cultures.

Norway and Finland have national systems, carefully curated by experts, to classify scholarly publication channels. The Norwegian system was developed between 2003 and 2004 (Sivertsen, 2016, 2018), whereas JUFO emerged from a project that began in 2010 (Pölönen, 2019). Although both systems predate DORA (2013), their use of discipline-based expert judgement and limited reliance on journal citation indicators align with certain responsible research assessment principles, provided that the classifications are used at an appropriate aggregate level. JUFO and the Norwegian Register are clear examples of the expert-based approach to classifying publication channels (Pölönen, 2019; Pölönen & Pylvänäinen, 2023).

In the systems of both Norway and Finland, experts who specialise in discipline-related areas assess journals, series, conferences, and publishers against established criteria before assigning specific publication channels at different levels. Different from citation indicators or general discovery services, these classifications are based on clear scholarly judgement regarding the quality and features of publication channels, rather than relying primarily on citation counts or automated web indexing.

- The JUFO classification designates publication channels as belonging to one of three levels: Level 1 (basic), Level 2 (leading), or Level 3 (highest-level). Level 0 comprises channels that have not been assessed or that do not meet the criteria for Level 1.
- Level 1 in the Norwegian Register indicates that a channel has been approved as an academic publication channel. Level 2 indicates that a channel has been approved at the highest level in its field, and Level 0 indicates that a channel is not approved because one or more of the applicable criteria have not been met or properly documented.

Such systems offer an alternative to citation-based journal indicators by taking into account field-specific judgments about the features and scholarly standing of publication venues. Yet merely replacing a citation-based indicator with one assigned by experts does not guarantee that the assessment methods align with the principles of responsible research assessment. Although both methods provide information about the publication venue, neither should be used in place of evaluating the content or the contribution of individual publications or researchers. The guidance on the responsible use of JUFO makes it clear that the classification was intended to assess the average quality of large groups of publications, not to judge the quality of individual works or to compare individual researchers.

Another difficulty arises when such classification systems are used in large-scale bibliographic workflows. Tasks such as systematic reviews, analyses of institutional publications, research monitoring, and certain evaluation procedures typically involve the systematic handling of hundreds or even thousands of bibliographic records. Although national classification systems generally offer searchable portals and sometimes provide registry records as structured downloads, the data are not always presented in a format that allows direct matching with the various bibliographic files that users might supply. Consequently, when applying registry classifications on a large scale, additional steps are needed, for example, cleaning bibliographic metadata, reconciling identifiers, normalising publication titles, handling historical versions of the classifications, and resolving ambiguous records or those that fail to match. Without an integrated alignment process, these tasks can require substantial manual effort and may therefore hinder the consistent application of national registry classifications across large datasets.

Table 1 presents a comparative taxonomy of selected publication-screening resources and approaches, categorised by their primary function, governance or data authority, degree of evaluative oversight, scalability of data access, and compatibility with RRA principles. Because the categories encompass different forms of

scholarly infrastructure—including indicators, indexing databases, discovery platforms, expert classification systems, and a metadata-alignment tool—the taxonomy compares their potential roles in publication screening rather than treating them as functionally equivalent alternatives.

Table 1. *Comparative taxonomy of scholarly publication-screening resources and approaches*

Resource or approach	Primary function and unit of analysis	Governance or data authority	Evaluative oversight and exposure to questionable outlets	Scalable data accessibility and computational usability	Compatibility with RRA principles
Journal citation indicators (e.g., JIF and CiteScore)	Measures citation performance aggregated at publication-channel level	Commercial providers, including Clarivate and Elsevier	Limited direct oversight: the indicator measures citation performance rather than independently evaluating editorial integrity; results may also be affected by disciplinary differences and strategic citation practices	Technically high but access-dependent: structured data and APIs may support automation, although access is generally subject to subscription, licensing, and reuse conditions	Limited when used as a surrogate: conflicts with RRA when journal-level scores substitute for assessment of individual outputs or researchers; may be used cautiously as contextual evidence
Curated indexing databases (e.g., Web of Science and Scopus)	Selects and indexes publications and publication channels	Commercial database providers and their editorial selection structures	Moderate to strong venue oversight: formal selection and review procedures provide safeguards, but questionable channels may remain indexed until re-evaluation or delisting	High for authorised users: structured exports and APIs support batch processing, subject to institutional access, licensing, and retrieval limits	Use-dependent: database inclusion can support discovery and verification but should not be treated as evidence of the quality of an individual publication
Broad scholarly discovery platforms (e.g., Google Scholar)	Automatically discovers and indexes scholarly and scholarly-adjacent documents	Proprietary search-engine algorithms	Limited qualitative oversight: broad automated coverage may include preprints, repositories, duplicate records, non-peer-reviewed documents, and material from questionable outlets	Low for reproducible batch processing: publicly searchable and broadly accessible, but lacks a supported general-purpose bulk API and restricts automated retrieval	Limited as an assessment mechanism: useful for discovery, but coverage or citation counts should not substitute for contextual and qualitative evaluation
National expert classification	Classifies journals, series,	National scholarly bodies,	Relatively strong venue-level oversight:	Moderate: portal searches and structured	Conditional: compatible with field-sensitive

systems (e.g., JUFO and the Norwegian Register)	conferences, and publishers at publication-channel level	including the Federation of Finnish Learned Societies and the Norwegian Directorate for Higher Education and Skills, supported by discipline-based expert structures	explicit criteria and expert review can identify unsuitable or questionable channels, although effectiveness depends on evidence availability, review cycles, and classification currency	downloads may be available, but applying the data to heterogeneous bibliographic files requires metadata reconciliation, historical-version management, and ambiguity resolution	aggregate assessment when used responsibly, but not appropriate as a proxy for the quality of individual publications or researchers
Metadata-driven registry alignment using SciFilter	Normalises bibliographic metadata and matches records to external registry classifications	Algorithmic processing based on periodically synchronised national registry data	Inherited and conditional oversight: conveys the classifications assigned by expert systems but does not independently evaluate journals or individual publications; results depend on registry coverage and currency	High for batch processing, with qualifications: automates matching and reporting across input files, while ambiguous, unmatched, or conflicting records may still require review	Conditional: can operationalise expert-curated classifications consistently at scale, but does not transform venue-level classifications into article-level quality assessments and must be used within the registries' responsible-use limitations

The analytical dimensions and categorical assignments in Table 1 form a heuristic taxonomy derived from the documented functions and characteristics of the respective resources and approaches. The qualitative assignments are not direct measures of system performance and should therefore be interpreted as transparent analytical classifications rather than empirically validated scores.

Evaluative oversight and exposure to questionable outlets hinge on the extent to which a resource employs explicit selection criteria, editorial review, or discipline-based expert assessment capable of identifying publication channels with inadequate scholarly or editorial practices. Metric-based assessment systems remain susceptible to strategic manipulation because using quantitative indicators as evaluative targets can incentivise behaviours intended to inflate apparent scholarly performance (Biagioli & Lippman, 2020). These vulnerabilities include citation manipulation and may intersect with broader threats to publication integrity, including questionable journals and paper-mill activity (Cortegiani et al., 2020; do Carmo et al., 2026). Automated discovery platforms provide limited qualitative oversight because their primary function is document discovery rather than publication-channel evaluation, allowing inadequately vetted materials to be discoverable alongside conventionally published research (Delgado López-Cózar et al., 2014). Curated commercial databases apply formal selection criteria but may not immediately identify compromised editorial practices. National classification systems provide comparatively strong venue-level oversight through defined criteria and expert review, although their effectiveness depends on available evidence, classification coverage, and the frequency of reassessment (Pölonen & Pylvänäinen, 2023; Sivertsen, 2016).

Scalable data accessibility and computational usability concern both the availability of structured data and the practical capacity to process records within reproducible, large-scale workflows. Bibliographic platforms vary substantially in their coverage, indexing models, and data-retrieval mechanisms (Martín-Martín et al., 2018). Commercial databases commonly provide structured exports or APIs, although access and reuse may be restricted by subscriptions, licences, and retrieval limits. Google Scholar provides broader and more inclusive coverage than curated bibliographic databases (Martín-Martín et al., 2018), but it does not offer a supported general-purpose bulk API, which limits its suitability for reproducible, automated screening. National classification systems may provide searchable portals and downloadable registry data, but applying these records to heterogeneous

bibliographic files still requires identifier reconciliation, metadata normalisation, historical version control, and the resolution of uncertain matches.

The extent to which an approach is compatible with the RRA principles depends on whether it enables a contextual, transparent, and appropriately qualitative evaluation, without using features of the publication channel as automatic stand-ins for the quality of a particular output or researcher. DORA (2013) warns against the use of journal-based metrics when assessing individuals, while the Leiden Manifesto (Hicks et al., 2015) stresses that quantitative indicators should support rather than replace expert judgement. The Agreement on Reforming Research Assessment likewise discourages inappropriate dependence on publication-based metrics and journal rankings (Arentoft et al., 2022). As a result, neither an indicator nor an expert-assigned rating at the publication channel level is inherently compatible with the RRA principles in all cases. National registry classifications and registry-alignment tools are only conditionally compatible, since their appropriateness depends on the unit in question, the purpose for which they are used, the level of transparency, and the consequences of their use.

Table 1 shows a three-way tension among evaluative oversight, the use of large amounts of data, and compatibility with the principles of responsible research assessment. Although citation indicators and commercial databases are generally suitable for large-scale computational processing, access to them may depend on proprietary infrastructure, and their indicators can contradict responsible research assessment when used as a substitute for evaluating individual research outputs. While broad discovery platforms offer extensive, easily searchable coverage, they provide only limited qualitative control over the publication channels and types of documents they include. National classification systems make use of expert judgement based on discipline and can provide reusable structured data, but further actions, such as metadata reconciliation and workflow integration, are needed if these classifications are to be applied consistently across a variety of bibliographic datasets.

Metadata-driven registry alignment provides a way of dealing with this integration challenge by separating the experts' task of classifying publication channels from the technical procedure of matching bibliographic records to these classifications. For instance, SciFilter standardises the input metadata and aligns the records with the data in the national registry by using identifiers and comparisons based on the title, the registry data being updated on a periodic basis. This method enables reproducible batch processing and ensures that there is traceability to the original registry classifications.

SciFilter does not evaluate the quality of publication channels or individual articles, and its outputs must not be taken as quality assessments at the article level. The results from SciFilter depend on the completeness of the registry, the up-to-dateness of the classifications, the quality of the input metadata, and the accuracy of the matching process. Since the registry is updated every quarter, recent changes might not be reflected immediately, and records that are ambiguous, in conflict, or that have not been matched may still need to be reviewed manually. The primary contribution of SciFilter is to reduce the effort required to apply expert-curated venue classifications consistently across large bibliographic datasets, not to replace expert judgement or eliminate the need for complete manual review.

SciFilter System Architecture

The architecture of the SciFilter system provides a structured approach to managing bibliographic data, implementing registry-alignment logic, and supporting scalable screening of publication sources. To maximise generalisability, this section emphasises the conceptual workflow for implementing metadata-driven screening. This workflow serves as a generalisable model for metadata-driven registry alignment, in which bibliographic metadata functions as the input layer, national registry systems provide the evaluative reference layer, and classification outputs represent operationalised evaluation decisions.

SciFilter functions as a metadata-driven screening system that enables evaluators to assess publication sources against established registry-based classification frameworks, specifically the JUFO and the Norwegian Register. The system supports bibliometric workflows, systematic literature reviews, and institutional evaluation processes requiring the rapid screening of large bibliographic datasets.

At a conceptual level, the system comprises three core operational components: (1) metadata ingestion and normalisation, (2) deterministic registry alignment and classification, and (3) structured output generation. Together, these components transform heterogeneous bibliographic inputs into standardised, registry-aligned classifications.

1. Metadata Ingestion and Normalisation

The system accepts bibliographic data via manual entry and bulk structured uploads (e.g., CSV, Excel, XML). Users provide core metadata fields, including authors' names, publication titles, publication years, source venues, and persistent identifiers (ISSNs or ISBNs). Upon ingestion, the pipeline parses and normalises the incoming

metadata to eliminate structural variances across export formats. Normalisation includes stripping string punctuation, converting text to lowercase, standardising hyphenation in ISSN/ISBN fields, and resolving publisher variant naming conventions that would otherwise cause alignment mismatches.

2. Deterministic Registry Alignment Engine

The core engine compares normalised metadata against a centrally maintained reference database containing JUFO and Norwegian classification snapshots updated quarterly. To ensure complete process auditability, SciFilter relies on a deterministic, rule-based execution waterfall instead of opaque probabilistic models or uninterpretable machine learning algorithms. Matching proceeds through a strict, sequential hierarchy:

1. **Exact Identifier Matching:** The engine first executes an exact string match on primary persistent identifiers (ISSN or eISSN for journals; ISBN for books and monographs).
2. **Exact Title Matching:** If an identifier is absent or unindexed, the system attempts an exact string match on the normalised publication source title.
3. **Fuzzy String Matching:** If exact matching fails, the engine executes fuzzy string alignment on normalised publication venue titles (comparing the ingested source title against official registry venue titles). The similarity score S is computed using normalised Levenshtein edit distance (Levenshtein, 1966):

$$S = 1 - \frac{d(a, b)}{\max(|a|, |b|)}$$

where $d(a, b)$ represents the minimum number of single-character edits (insertions, deletions, or substitutions) required to transform input title a into registry title b . The engine accepts candidate matches only when the similarity score satisfies the empirically calibrated threshold of $S \geq 0.90$.

If an identifier links to two distinct registry records, or if fuzzy matching yields multiple candidates that exceed the similarity threshold, the system flags the entry for manual review to prevent false-positive classifications. If no candidate meets the threshold at any stage, the source receives a "Not Indexed" label.

3. Structured Output Generation and Auditability

The system generates enriched export datasets that display the indexed status and the exact registry tier (e.g., Levels 0–3) for each venue. Crucially, SciFilter operationalises human-curated, expert-based classification frameworks through automated metadata alignment. Because every decision follows a deterministic rule, each result maintains complete lineage tracing, enabling researchers and administrators to audit the precise factor that produced a given classification while eliminating manual web portal queries.

Methods

Materials

Regarding the Evaluation Benchmark Dataset, the validation dataset comprised $N = 500$ unique scholarly publication venues, structured with a 1:1 balance between indexed sources ($n = 250$) and non-indexed sources ($n = 250$). Regarding Data Sourcing and Metadata, part of the dataset was retrieved via the Open Academy Journal Index (OAJI) website¹ (particularly metadata for journals retrievable via Google Scholar), with additional sources drawn from Web of Science² and Scopus³. Extracted metadata fields included ISSNs, ISBNs, authors, titles, and publication years, which were collected and verified. Regarding Sample Diversity and Range, the resulting dataset represents a heterogeneous, interdisciplinary sample spanning education, information systems, learning sciences, management, and STEM-related fields, with publication years ranging from 1978 to 2025. Across formats, the 500 curated sources comprise journals, conference proceedings, book chapters, and one arXiv

¹ <https://oaji.net/journals-list.html?discipline>

² <https://www.webofscience.com/>

³ <https://www.scopus.com/home.uri>

preprint. In terms of Scope and Limitations, we do not intend the dataset to be statistically representative of the global publication landscape; rather, it serves as a controlled evaluation set to assess agreement between human and automated registry alignment. While this design enables systematic comparison, it may limit generalisability to broader real-world datasets, particularly those with noisier or incomplete metadata. Regarding National Quality Registries, ground-truth classification ratings were compiled from local database copies of two European field-normalised quality registries:

- *JUFO*: Levels 0-3;
- *Norwegian Register*: Levels 0-2.

The full composition and stratification of the 500-source benchmark dataset across both rating systems is detailed in Table 2.

Table 2. *Breakdown of the 500 Curated Sources.*

Source type	Numbers included
Ranked Journals	273
Ranked Conference Proceedings	11
Ranked book chapters	5
Not Ranked Conference Proceedings	12
Not Ranked book chapters	37
Journals flagged as potential problematic in prior literature or public listings	161
Arxiv preprint	1
Total	500

*Note: 'Ranked' and 'Not Ranked' describe source type categories eligible for registry classification, not indexing outcome. Of the 289 ranked-type sources, 39 received a Level 0 (not indexed) classification; combined with the 211 not-ranked-type sources, this yields the reported 1:1 balance of 250 indexed and 250 non-indexed sources.

Procedure

The empirical evaluation of SciFilter was governed by a structured, three-phase experimental protocol designed to evaluate the system's automated indexing verification accuracy, multi-class error distribution, and operational applications. As depicted in Figure 2, the methodology maps directly to the study's primary research questions: Phase I evaluates the reliability of the benchmark dataset to test RQ1, Phase II applies the core screening engine central to RQ3, and Phase III performs the statistical alignment and error analysis serving as the primary empirical evidence for RQ2.

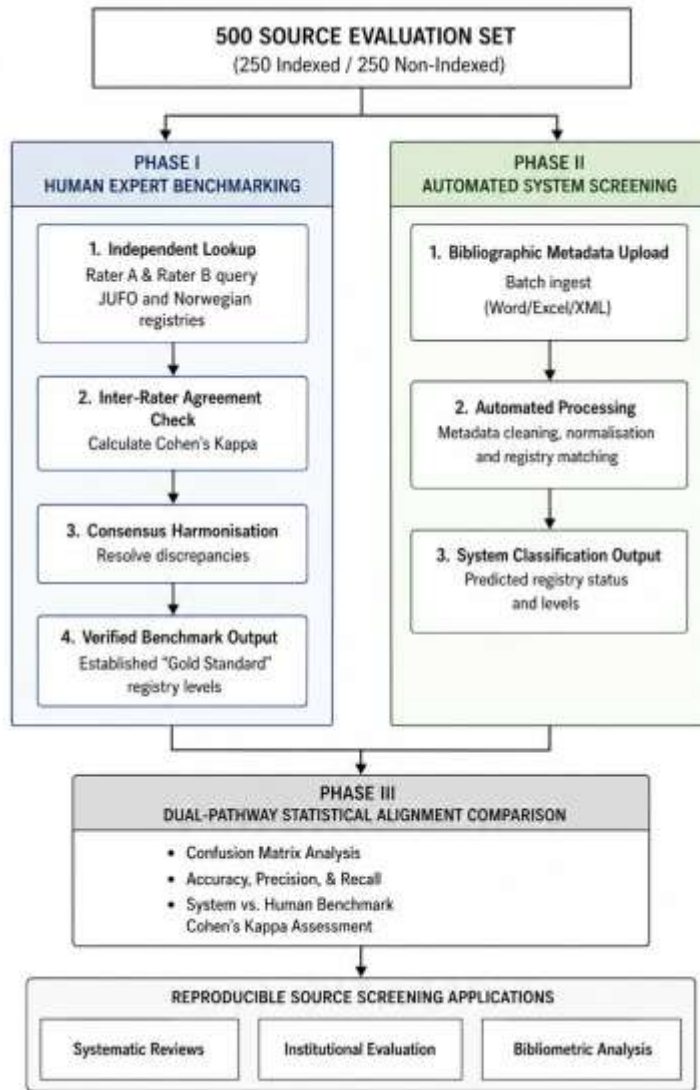


Fig. 2 Three-phase experimental workflow for SciFilter validation, detailing Phase I (Human Expert Benchmarking), Phase II (Automated System Screening), and Phase III (Dual-Pathway Statistical Alignment).

Phase I: Human Expert Benchmarking (RQ1 Validation)

To establish an authoritative ground-truth baseline for system validation (RQ1), we executed a four-stage human evaluation protocol across 500 scholarly publication venues:

1. *Independent Lookup*: Two researchers (Rater A and Rater B) independently queried the online portals of both the JUFO and the Norwegian Register for each venue. Each venue was assigned its respective numerical tier or flagged as non-indexed.
2. *Inter-Rater Agreement Check*: To verify that the ground-truth assignment process was objective and reproducible, initial blind ratings between Rater A and Rater B were cross-tabulated. Inter-rater agreement was quantified using Cohen (1960)'s Kappa (described fully in Phase III).
3. *Consensus Harmonisation*: The few instances of rating discrepancies (primarily stemming from historical registry-level updates or ambiguous volume series) were flagged and jointly reviewed during a harmonisation session, producing a single, corrected reference set used specifically for precision, recall, and accuracy testing (Phase III).
4. *Verified Benchmark Output*: This harmonised "Gold Standard" served as the benchmark for classification-accuracy metrics. Separately, Rater A's and Rater B's original, unharmonised independent judgments were retained to assess inter-rater and human-machine reliability (Cohen's and

Fleiss' Kappa), so that reliability estimates would not be inflated by post-hoc correction of the very disagreements they are meant to measure.

Phase II: Automated System Screening (RQ3 Task Application)

Phase II evaluated the practical operational capacity of SciFilter (RQ3) by simulating real-world batch processing workflows:

1. *Bibliographic Metadata Upload*: The 500 benchmark venues were converted into standardised tabular data formats (Excel, CSV, XML) containing metadata fields for source titles, print ISSNs, electronic ISSNs, and ISBNs. This file was ingested into SciFilter via its bulk processing module.
2. *Automated Processing and Matching Cascade*: The core screening engine parsed the incoming metadata and executed a deterministic, three-tiered matching cascade against internally mirrored database tables of the JUFO and Norwegian registries:
 - i. *Tier 1 (Exact Identifier Matching)*: Direct string comparison against standard alphanumeric identifiers (print-ISSN, electronic-ISSN, and ISBN).
 - ii. *Tier 2 (String Normalisation and Exact Title Match)*: For entries lacking valid numerical identifiers, titles were pre-processed (lowercasing, punctuation stripping, and whitespace collapse) and matched against normalised registry titles.
 - iii. *Tier 3 (Fuzzy Title Matching)*: Unmatched variant titles were passed to a Levenshtein distance similarity algorithm. To eliminate false positives, a strict similarity threshold (≥ 0.90 threshold) was enforced.
3. *System Classification Output*: SciFilter automatically logged each source's predicted index status (indexed against non-indexed) and assigned field-normalised quality levels.

Phase III: Dual-Pathway Statistical Alignment & Error Analysis (RQ2 Evidence)

To generate primary empirical evidence regarding classification accuracy and error behaviour (RQ2), system-predicted classifications were cross-tabulated against the Phase I human expert ground truth:

Binary Classification Metrics: System outcomes were mapped into a 2×2 confusion matrix comprising True Positives (TP), False Positives (FP), False Negatives (FN), and True Negatives (TN). From this matrix, standard statistical performance metrics were calculated:

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Statistical Reliability Assessment: To measure overall statistical alignment between automated predictions and expert consensus beyond chance, Cohen's Kappa (κ) was computed:

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

where P_o represents the observed proportion of agreement between SciFilter and the human benchmark, and P_e represents the expected proportion of chance agreement. Because Cohen's Kappa is defined only for two raters at a time, it was used to assess pairwise agreement (between Rater A and Rater B, Rater A and SciFilter, and Rater B and SciFilter), while Fleiss (1971)' Kappa — which generalises the same logic to more than two raters — was used to assess agreement across all three raters (Rater A, Rater B, and SciFilter) simultaneously.

The Fleiss' Kappa is obtained as follows:

For N subjects, each rated by n raters into one of k categories:

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$$

Where:

Per-subject agreement P_i , for subject i , letting n_{ij} = number of raters who assigned subject i to category j :

$$P_i = \frac{1}{n(n-1)} \left(\sum_{j=1}^k n_{ij}^2 - n \right)$$

Mean agreement across all subjects:

$$\bar{P} = \frac{1}{N} \sum_{i=1}^N P_i$$

Category marginal proportion p_j , for category j , across all $N \times n$ ratings:

$$p_j = \frac{1}{Nn} \sum_{i=1}^N n_{ij}$$

Expected agreement by chance:

$$\bar{P}_e = \sum_{j=1}^k p_j^2$$

Multi-Class Tier Error Analysis: Beyond binary detection (indexed against non-indexed), classification error distributions across multi-tiered quality levels (JUFO Levels 0–3 and Norwegian Levels 0–2) were isolated to determine whether system errors were catastrophic (e.g., Level 0 misclassified as Level 3) or restricted to adjacent levels (e.g., Level 1 against Level 2). Detailed procedural protocols for data preparation and manual verification are provided in Appendix A.

AI Use

During manuscript preparation, AI assistants (Claude, Anthropic, and Google Gemini) helped reorganise and consolidate several already-stated points into clearer locations in the text and flagged redundant or inconsistent passages for review. ChatGPT (OpenAI) supported the evaluation study design through iterative consultations. No decisions, judgments, or conclusions were delegated to these tools, and all AI-assisted content was reviewed and revised by the authors before inclusion. The authors independently verified, reviewed, and approved the final study design, data interpretation, and manuscript content, in line with Springer Nature's editorial policies on the responsible use of AI in scholarly communication and with COPE guidance on AI in publishing.

Results

In this section, we present the findings from the evaluation of SciFilter, focusing on the degree of agreement between system-generated classifications and expert-curated registry assignments. The results are interpreted in terms of consistency, reliability, and the practical implications of metadata-driven registry alignment for large-scale screening tasks. The evaluation comprises three components: inter-rater agreement between human

evaluators, system performance relative to the benchmark dataset, and detailed analysis of classification outcomes using confusion matrices.

Inter-rater Results and Ground-Truth Reliability (Phase I/RQ1)

To evaluate the baseline reliability of the ground-truth benchmark, two researchers independently classified the 500 curated sources across three dimensions: the source's JUFO level, the corresponding Norwegian level, and the binary indexing status ("Indexed" or "Not Indexed").

Table 3. *Rater A against Rater B (human-human reliability)*

	B: Indexed	B: Not Indexed
A: Indexed	240	10
A: Not Indexed	2	248

For the binary indexing status, Raters A and B both agreed that 240 sources were indexed and 248 were not, yielding an observed agreement of $P_o = \frac{488}{500} = 0.976$. The expected agreement was calculated as:

$$P_e = \left(\frac{250}{500} \times \frac{242}{500}\right) + \left(\frac{250}{500} \times \frac{258}{500}\right) = 0.242 + 0.258 = 0.5$$

Calculating Cohen's Kappa yielded:

$$k = \frac{0.976 - 0.5}{1 - 0.5} = \frac{0.476}{0.5} = 0.95$$

This result falls within the 0.81 – 1.00 range, indicating near-perfect agreement between two human expert raters according to standard benchmark interpretations (Table 4).

Table 4. *Cohen's Kappa Values and Interpretation*

Kappa Value	Interpretation
≤ 0	No agreement
0.01–0.20	Slight agreement
0.21–0.40	Fair agreement
0.41–0.60	Moderate agreement
0.61–0.80	Substantial agreement
0.81–1.00	Almost perfect agreement

For multi-class-level rankings, Raters A and B agreed on 241 sources correctly ranked at the JUFO level and 244 at the Norwegian level, resulting in an observed agreement of $P_o = \frac{485}{500} = 0.97$. The expected agreement was calculated as:

$$P_e = \left(\frac{241}{500} \times \frac{244}{500}\right) + \left(\frac{259}{500} \times \frac{256}{500}\right) = 0.235 + 0.265 = 0.5$$

The resulting Cohen's Kappa was $k = \frac{0.97 - 0.5}{1 - 0.5} = 0.94$, which also falls within the 0.81 – 1.00 range, confirming almost perfect agreement between the raters.

A qualitative review of the locations where discrepancies occurred revealed that they were primarily due to human error in accurately recording observations rather than fundamental disagreement over registry rules. For example, in three instances, both curators correctly entered the source ranking levels, but Curator B incorrectly classified the status as "Not Indexed". Similarly, Curator A provided incorrect JUFO and Norwegian-level rankings for two sources. After reviewing and correcting these recording discrepancies, the final reference benchmark dataset was

established for precision, recall, and accuracy testing (Phase III). To avoid inflating reliability estimates through post hoc correction of the very disagreements they are meant to capture, the original, unharmonised judgments of Rater A and Rater B — prior to consensus review — were retained separately for inter-rater and human-machine reliability analyses. Three-rater reliability (Rater A, Rater B, and SciFilter) was then assessed using Fleiss' Kappa on the unharmonised data, as described below.

Fleiss' Kappa, computed over all 500 subjects $\times$ 3 raters $\times$ 2 categories:

Category marginals across all 1,500 individual ratings: 740 "indexed" (49.33%), 760 "not indexed" (50.67%)

$$\bar{P} = 0.9707$$

$$\bar{P}_e = 0.4933^2 + 0.5067^2 = 0.2434 + 0.2567 = 0.5001$$

$$\text{Fleiss' } \kappa = (0.9707 - 0.5001) / (1 - 0.5001) = 0.9413 (\sim 0.94)$$

System Testing and Comparison Results (Phase II & III / RQ2 & RQ3)

During automated system testing against the verified benchmark, 233 sources were correctly identified as indexed with the correct rank (*TP*), 12 sources were wrongly identified as indexed or misranked (*FP*), 10 indexed sources were not identified or misclassified (*FN*), and 245 sources were correctly identified as not indexed (*TN*). Evaluating these raw classification counts across standard system reliability metrics yielded a Precision of 0.95 ($\frac{233}{245}$), a Recall of 0.96 ($\frac{233}{243}$), and an Overall System Accuracy of 0.96 ($\frac{478}{500}$).

In addition to precision, recall, and accuracy, inter-rater reliability between SciFilter and each human rater was assessed using Cohen's Kappa, and reliability across all three raters (Rater A, Rater B, and SciFilter) jointly was assessed using Fleiss' Kappa.

Table 5. *Rater A against SciFilter*

	SciFilter: Indexed	SciFilter: Not Indexed
A: Indexed	242	8
A: Not Indexed	6	244

$$P_o = (242 + 244) / 500 = 0.972$$

$$P_e = (250/500 \times 248/500) + (250/500 \times 252/500) = 0.248 + 0.252 = 0.500$$

$$\text{Cohen's } \kappa = (0.972 - 0.500) / (1 - 0.500) = 0.94$$

Table 6. *Rater B against SciFilter*

	SciFilter: Indexed	SciFilter: Not Indexed
B: Indexed	236	6
B: Not Indexed	12	246

$$P_o = (236 + 246) / 500 = 0.964$$

$$P_e = (242/500 \times 248/500) + (258/500 \times 252/500) = 0.2401 + 0.2601 = 0.5001$$

$$\text{Cohen's } \kappa = (0.964 - 0.5001) / (1 - 0.5001) = 0.93$$

Confusion Matrix Analysis & Level Discrepancies (RQ2)

To assess error severity across multi-tiered classifications, system predictions were compared with individual registry ranks using confusion matrix visualisations.

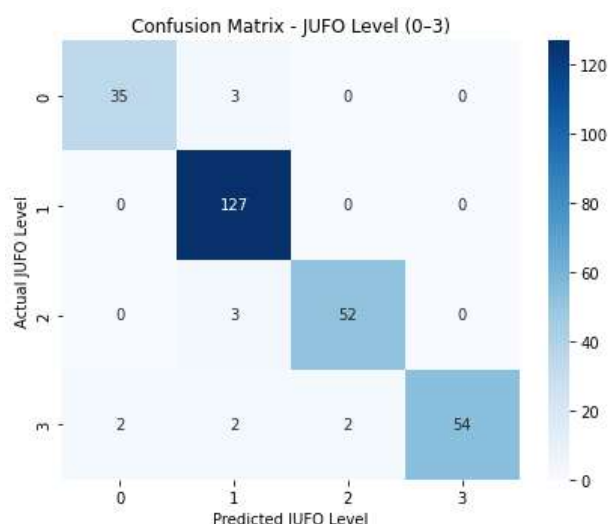
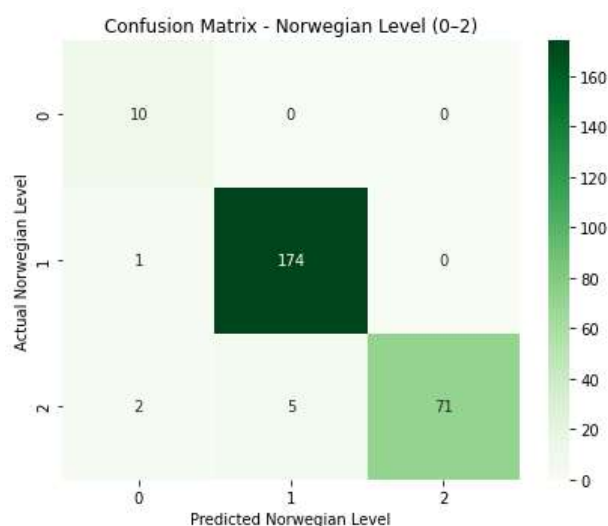


Fig. 3⁴ Confusion matrix displaying actual against predicted classification counts across JUFO Levels 0–3. A higher degree of concordance is indicated by more pronounced shading along the diagonal, indicating greater predictive accuracy.

The JUFO confusion matrix (Figure 3), covering the 280 sources with an assigned JUFO level, indicates that SciFilter correctly classified approximately 95.7% of these sources. Most classification errors occurred between adjacent levels, such as Level 0 misclassified as Level 1 and Level 2 misclassified as Level 1. Across the specific tier cells:

- Level 0: 35 sources were correctly predicted as Level 0, while 3 sources were misclassified as Level 1.
- Level 1: 127 sources were correctly predicted as Level 1, with zero false classifications into other tiers.
- Level 2: 52 sources were correctly predicted as Level 2, while 3 sources were misclassified as Level 1.
- Level 3: 54 sources were correctly predicted as Level 3, while 2 sources were misclassified as Level 0, 2 as Level 1, and 2 as Level 2.



⁴ Sources absent from a given registry are not included in that registry's confusion matrix, as no tier exists to predict against.

Fig. 4⁵ Confusion matrix displaying actual against predicted classification counts across Norwegian Register Levels 0–2. A higher degree of concordance is indicated by more pronounced shading along the diagonal, indicating greater predictive accuracy.

Figure 4 presents the corresponding confusion matrix for the 263 sources assigned a Norwegian Register level (Levels 0–2). Across specific tier cells:

- Level 0 (Unrated/Non-Indexed): 10 sources were correctly predicted as Level 0, with zero misclassifications into higher tiers.
- Level 1 (Standard Academic): 174 sources were correctly predicted as Level 1, while 1 source was misclassified as Level 0.
- Level 2 (Top-Level Academic): 71 sources were correctly predicted as Level 2, while 2 sources were misclassified as Level 0 and 5 sources were misclassified as Level 1.

Overall, across both registries, the system exhibited high exact-tier alignment. Most classification discrepancies occurred at adjacent tier boundaries, consistent with expected boundary ambiguity between similarly rated sources; a smaller number involved larger tier discrepancies — most notably four JUFO Level 3 sources misclassified as Level 0 or Level 1 — which represent a more consequential class of error and merit closer review in future work.

Discussion

The evaluation of SciFilter demonstrates a high level of statistical agreement between metadata-driven classifications and expert-curated registry assignments across both indexed and non-indexed publication sources ($k=0.94$; overall system accuracy = 96%). As opposed to indicating independent validation of publication quality, these findings highlight the extent to which registry-based evaluation logic can be operationalised and reproduced through structured metadata processing. Importantly, instead of replacing expert-based evaluation, registry-aligned screening operationalises it, enabling the consistent, scalable application of frameworks such as JUFO, which were designed for manual, individual-level lookup. This addresses a key limitation in current research evaluation practices, where expert-based systems exist but are difficult to implement in high-volume contexts. This result is significant for bibliometric workflows, where consistency and reproducibility of screening decisions are critical, particularly for large-scale tasks such as systematic reviews, institutional reporting, and research evaluation.

The high levels of agreement observed in both binary classification (0.95 precision, 0.96 recall) and multi-level classification (95.7% exact-tier alignment for JUFO; 97.0% for the Norwegian Register) suggest that metadata-driven registry alignment can serve as a reliable first-pass screening mechanism. The confusion matrix analyses (Figures 3 and 4) indicate that most classification errors occur between adjacent levels rather than between distant categories. This pattern suggests that the system effectively captures the general structure of registry-based classifications, with discrepancies arising primarily at classification boundaries rather than from fundamental misinterpretation. These findings contribute to ongoing discussions in bibliometrics about the operationalisation of evaluation frameworks. While prior work has focused on citation-based indicators (Garfield, 1972; Hirsch, 2005) and their limitations (Seglen, 1997; Fire & Guestrin, 2019), this study illustrates how registry-based systems can be translated into scalable, metadata-driven processes without relying on citation data.

In contrast to citation-based metrics, which may reflect dissemination and visibility rather than methodological rigour (Moed, 2005; Pastorino et al., 2016; Thelwall et al., 2023a; Thelwall et al., 2023b), registry-based classifications incorporate expert judgement and field-sensitive criteria. By operationalising these classifications through automated screening, SciFilter offers a complementary approach that aligns with the principles of responsible research assessment (Hicks et al., 2015). Our study also complements work such as Nguyen et al. (2022), which deploy a data quality appraisal framework, particularly for bibliographic data. Recent bibliometric work has begun using these expert-curated registries as external validation benchmarks for new citation-based metrics: Barnes et al. (2026) find that journals ranked highly by JUFO and the Norwegian Register tend to receive citations from more distant, less self-referential sources — treating registry classifications as a credible reference point against which a novel metric's validity can be judged. From a practical perspective, the ability to process large bibliographic datasets and produce reproducible screening outputs positions metadata-

⁵ Sources absent from a given registry are not included in that registry's confusion matrix, as no tier exists to predict against.

driven registry alignment as a useful component of research evaluation pipelines. This is particularly relevant for early-stage researchers, librarians, and institutional research offices that require efficient and transparent methods for assessing publication sources.

These findings have implications for institutional research evaluation systems, particularly in contexts where registry-based classifications are used to allocate funding, assess performance, or set publication benchmarks. By enabling the reproducible application of such systems, metadata-driven alignment may contribute to greater transparency and consistency in evaluation processes.

Limitations and Future Work

At the same time, four limitations must be considered:

- **Registry Coverage Boundaries:** The evaluation relies on JUFO and the Norwegian Register as reference frameworks, both of which have limited coverage and exclude some legitimate publication venues. As a result, absence from these registries should not be taken as an indicator of low quality, but rather as a signal for further scrutiny.
- **Controlled Benchmark Composition:** The dataset used in this study was manually curated and balanced across indexed and non-indexed sources (N=500, 1:1 ratio). While this enables controlled evaluation, it may not fully reflect the distribution and variability of real-world bibliographic datasets, particularly those with incomplete or noisy metadata.
- **Circular Baseline Evaluation:** The evaluation design focuses on agreement between system outputs and expert-informed registry classifications, both of which are based on the same underlying reference frameworks.
- **Metadata Integrity Dependencies:** The system relies on accurate, standardised metadata (such as clean ISSN/ISBN strings and uncorrupted title fields) for match effectively. Variations in naming conventions, missing identifiers, or inconsistent formatting may affect classification outcomes.

Future work should assess the robustness of metadata-driven screening under conditions of incomplete or noisy data, which are common in real-world applications. Finally, the current implementation is primarily focused on Scandinavian and European registry systems. Expanding the approach to include additional classification frameworks and global indexing systems would improve generalisability and applicability across diverse research contexts.

Overall, these findings suggest that metadata-driven registry alignment offers a viable pathway to scale expert-based evaluation systems, while also highlighting the importance of combining automated tools with contextual interpretation and human judgement in research assessment.

Conclusion

This study demonstrates that registry-based publication classification systems can be put into practice using metadata-driven methods, enabling a scalable and reproducible screening of scholarly sources without the need to develop new metrics-based indicators. In the context of the current debates in metascience concerning the difficulties of peer review and research assessment (Howe, 2025), there is a strong need for filtering systems that are both efficient and in line with the principles of responsible research evaluation (Hicks et al., 2015).

This study shows that when SciFilter is assessed using a benchmark consisting of 500 sources, it can reliably replicate the screening decisions made by experts (RQ1). The system achieved a high level of operational accuracy (0.96), precision (0.95), and recall (0.96), resulting in strong statistical agreement with human curators ($k = 0.94$). Moreover, the multi-class error analysis revealed that most discrepancies were limited to adjacent quality levels, though a smaller number of larger tier jumps (e.g., Level 3 sources misclassified as unindexed) indicate that boundary and outlier cases still warrant human review. This confirms that automated pipelines can act as a reliable first-stage filter, not a final arbiter, in quality-control procedures (RQ2).

Conclusively, metadata-driven alignment provides a practical and theoretically sound means of scaling expert-based evaluation models across a wide range of high-volume bibliometric processes—such as systematic reviews, institutional reporting, and the allocation of resources—where manual lookup is not feasible (RQ3). Instead of replacing human review or acting as a final guarantee of legitimacy, SciFilter shows how automated tools can be combined with contextual human judgement to enhance transparency, consistency, and scalability in relation to modern research assessment.

Funding

This study was co-funded by the European Union and the Estonian Ministry of Education and Research through the TEM-TA26 project: Digital Transformation Through Lifelong Learning.

Data Availability Statement

The data supporting the findings of this study are openly available on the Open Science Network platform.

<https://doi.org/10.17605/OSF.IO/V53WU>

The tool SciFilter is publicly available at <https://www.scifilter.eu>.

Conflict of Interest Declaration

The authors declare no conflicts of interest.

Author contributions

A.A.O. conceived and designed the study, developed the methodology, conducted the formal analysis, and wrote the original draft. M.B. supervised the study and reviewed and edited the manuscript. A.E.O. contributed to the methodology and formal analysis and assisted in writing the manuscript. All authors reviewed the manuscript.

References

- Arentoft, M., Berghmans, S., Borrell-Damian, L., Bottaro, S., Faure, J.-E., Gaillard, V., Glinos, K., Lopez Albacete, J., Morais, R., Morris, J., Schiltz, M., & Stroobants, K. (2022). *Agreement on reforming research assessment* (Version 1). Zenodo. <https://doi.org/10.5281/zenodo.13480728>
- Barnes, K., Bannour, A., Török, L., & Molontay, R. (2026). Citation distance matters: Towards a new metric for evaluating journal impact. *Scientometrics*, 131, 1505–1524. <https://doi.org/10.1007/s11192-026-05586-1>
- Biagioli, M., & Lippman, A. (2020). Introduction: Metrics and the new ecologies of academic misconduct. In M. Biagioli & A. Lippman (Eds.), *Gaming the metrics: Misconduct and manipulation in academic research* (pp. 1–23). The MIT Press. <https://doi.org/10.7551/mitpress/11087.003.0001>
- Bollen, J., Van de Sompel, H., Hagberg, A., & Chute, R. (2009). A principal component analysis of 39 scientific impact measures. *PLOS ONE*, 4(6), e6022. <https://doi.org/10.1371/journal.pone.0006022>
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46. <https://doi.org/10.1177/001316446002000104>
- Cortegiani, A., Manca, A., Lalu, M., & Moher, D. (2020). Inclusion of predatory journals in Scopus is inflating scholars' metrics and advancing careers. *International Journal of Public Health*, 65(1), 3–4. <https://doi.org/10.1007/s00038-019-01318-w>
- Cronin, B., & Sugimoto, C. R. (Eds.). (2014). *Beyond bibliometrics: Harnessing multidimensional indicators of performance*. The MIT Press.
- Delgado López-Cózar, E., Robinson-García, N., & Torres-Salinas, D. (2014). The Google Scholar experiment: How to index false papers and manipulate bibliometric indicators. *Journal of the American Society for Information Science and Technology*, 65(3), 446–454. <https://doi.org/10.1002/asi.23056>

- do Carmo, F. A., Rezende, S. O., Paxiúba, C. M. C., & Lobato, F. M. F. (2026). Predatory publishing in the academic landscape: Researchers' awareness and vulnerabilities. *Scientometrics*, 131, 3029–3054. <https://doi.org/10.1007/s11192-026-05614-0>
- Fire, M., & Guestrin, C. (2019). Over-optimization of academic publishing metrics: Observing Goodhart's law in action. *GigaScience*, 8(6), giz053. <https://doi.org/10.1093/gigascience/giz053>
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 378–382.
- Garfield, E. (1972). Citation analysis as a tool in journal evaluation: Journals can be ranked by frequency and impact of citations for science policy studies. *Science*, 178(4060), 471–479. <https://doi.org/10.1126/science.178.4060.471>
- Hicks, D., Breitzman, A., Hamilton, K., & Narin, F. (2000). Research excellence and patented innovation. *Science and Public Policy*, 27(5), 310–320. <https://doi.org/10.3152/147154300781781805>
- Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature*, 520(7548), 429–431. <https://doi.org/10.1038/520429a>
- Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102(46), 16569–16572. <https://doi.org/10.1073/pnas.0507655102>
- Hogg, I. H., & Smith, J. R. (1959). Information and literature use in a research and development organization. In *Proceedings of the International Conference on Scientific Information* (Vol. 1, pp. 131–162). National Academy of Sciences–National Research Council.
- Hojnik, J. (2021). What shall I compare thee to? Legal journals, impact, citation and peer rankings. *Legal Studies*, 41(2), 252–275. <https://doi.org/10.1017/lst.2020.43>
- Howe, N. P. (2025). Metascience can improve science—but it must be useful to society, too. *Nature*, 643(8071), 304. <https://doi.org/10.1038/d41586-025-02065-0>
- Kitchenham, B., & Charters, S. (2007). *Guidelines for performing systematic literature reviews in software engineering* (Technical Report EBSE-2007-01). Keele University and Durham University.
- Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8), 707–710.
- Martín-Martín, A., Orduna-Malea, E., Thelwall, M., & Delgado López-Cózar, E. (2018). Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. *Journal of Informetrics*, 12(4), 1160–1177. <https://doi.org/10.1016/j.joi.2018.09.002>
- McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, 22(3), 276–282.
- Moed, H. F. (2005). *Citation analysis in research evaluation*. Springer. <https://doi.org/10.1007/1-4020-3714-7>
- Narin, F., Hamilton, K. S., & Olivastro, D. (1997). The increasing linkage between U.S. technology and public science. *Research Policy*, 26(3), 317–330. [https://doi.org/10.1016/S0048-7333\(97\)00013-9](https://doi.org/10.1016/S0048-7333(97)00013-9)
- Nguyen, B. X., Luczak-Roesch, M., Dinneen, J. D., & Larivière, V. (2022). Assessing the quality of bibliographic data sources for measuring international research collaboration. *Quantitative Science Studies*, 3(3), 529–559. https://doi.org/10.1162/qss_a_00211

Pastorino, R., Milovanovic, S., Stojanovic, J., Efremov, L., Amore, R., & Boccia, S. (2016). Quality assessment of studies published in open access and subscription journals: Results of a systematic evaluation. *PLOS ONE*, 11(5), e0154217. <https://doi.org/10.1371/journal.pone.0154217>

Pölönen, J. (2019). *What is Publication Forum?* Federation of Finnish Learned Societies.

Pölönen, J., & Pylvänäinen, E. (2023). *Publication Forum review of ratings in 2022*. Federation of Finnish Learned Societies. <https://doi.org/10.23847/tsv.618>

San Francisco Declaration on Research Assessment. (2013). *San Francisco Declaration on Research Assessment (DORA)*. <https://sfdora.org/read/>

Seglen, P. O. (1997). Why the impact factor of journals should not be used for evaluating research. *BMJ*, 314(7079), 497–502. <https://doi.org/10.1136/bmj.314.7079.497>

Sivertsen, G. (2016). Publication-based funding: The Norwegian model. In M. Ochsner, S. E. Hug, & H.-D. Daniel (Eds.), *Research assessment in the humanities: Towards criteria and procedures* (pp. 79–90). Springer. https://doi.org/10.1007/978-3-319-29016-4_7

Sivertsen, G. (2018). The Norwegian model in Norway. *Journal of Data and Information Science*, 3(4), 3–19. <https://doi.org/10.2478/jdis-2018-0017>

Thelwall, M., Kousha, K., Abdoli, M., Stuart, E., Makita, M., Wilson, P., & Levitt, J. (2023a). Why are coauthored academic articles more cited: Higher quality or larger audience? *Journal of the Association for Information Science and Technology*, 74(7), 791–810. <https://doi.org/10.1002/asi.24755>

Thelwall, M., Kousha, K., Stuart, E., Makita, M., Abdoli, M., Wilson, P., & Levitt, J. (2023b). In which fields are citation indicators of research quality? *Journal of the Association for Information Science and Technology*, 74(8), 941–953. <https://doi.org/10.1002/asi.24767>

Appendix A: Data Preparation and Manual Screening Details

This appendix outlines the technical procedures underpinning the three-phase evaluation described in the Methods section. These details are provided to support transparency and reproducibility while maintaining readability in the main text.

After compiling the curated list in Google Sheets and Microsoft Excel, we created a helper column and used the COUNTIF function (for example, =IF(COUNTIF(A:A, A2) > 1, "Duplicate", "Unique")) along with sort techniques in Google Sheets to flag each record as either “unique” or duplicated. This ensured that only one record per source was retained and that duplicate sources were eliminated, resulting in the desired 500 curated records.

After curating 500 records, the first and third authors independently searched the JUFO portal for sources using source names, ISSNs, or ISBNs. The rankings of each source according to JUFO and Norwegian standards were entered into the Google Sheet in the respective columns, and the status was marked as “Indexed” or “Not indexed” based on our defined parameters. A reliability test using Cohen’s kappa was then performed. Following this test, a further system test in SciFilter was conducted, and the results were compared with those from the JUFO portal. The evaluation was conducted in three phases.

Phase I: Manual Benchmarking

A benchmark reference dataset, based on expert consensus within a defined registry framework, was created to serve as the baseline for comparison in Phase II, when SciFilter’s output was evaluated against it.

We provide the following instructions to verify the indexing status. For each publication source (not the article itself), researchers were to search the JUFO portal and Norwegian Register (both classifications are on the portal), following the instructions below:

1. JUFO portal - <https://jfp.csc.fi/jufportal>
2. Search by:
 - Source name
 - ISSN (if available)
 - ISBN (if available)
3. Record:
 - JUFO Level (0–3)
 - Norwegian Level (if applicable) 0-2
 - Not Found
 - In the spreadsheet, enter JUFO and Norwegian levels. If source, ISSN, or ISBN is missing, put “Not Found” (JUFO), “N/A” (Norwegian), and “Not Indexed” (Status). Enter “Not Indexed” if level 0; otherwise, “Indexed” for levels 1–3.

Higher levels show greater recognition: JUFO 1–3, Norwegian 1–2, with level 2 being more selective. “Indexed” refers to registry inclusion, not quality. We assessed interrater agreement using Cohen’s Kappa to compare the two researchers’ assessments of the curated list against the JUFO portal records. Cohen’s Kappa is a statistical measure that evaluates inter-rater reliability — how much two raters agree beyond chance when categorising items. When two or more curators classify journals as “*Indexed*” or “*Not Indexed*”, Cohen’s Kappa helps determine if their agreement is consistent and trustworthy, not just due to luck. The interpretation of the Cohen’s Kappa ranges in Table 3 follows commonly cited benchmarks in the inter-rater reliability literature (see, for example, McHugh, 2012).

Phase II: System Testing

We benchmarked the curated list of 500 sources with the SciFilter engine by uploading it to the system. We ran the reference benchmark dataset through SciFilter and recorded system outputs (including "indexed" and "not indexed" cases).

Phase III: Comparison

We matched system-generated labels with expert-assigned registry labels, classifying results as TP (True Positive), FP (False Positive), FN (False Negative), or TN (True Negative), and used a confusion matrix to visualise errors. TP, FP, FN, and TN indicate correctly or incorrectly indexed sources. We computed Precision $[TP/(TP+FP)]$, Recall $[TP/(TP+FN)]$, and Accuracy $[(TP+TN)/(TP+FP+FN+TN)]$. A Python script using Pandas, seaborn, Matplotlib.pyplot, and sklearn.metrics visualised the results with confusion matrices.