From Manual Curation to Automated Extraction: The Economic and Scholarly Case for Multilingual Metadata Systems in Academic Publishing

Rashid Turgunbaev
ORCID: 0000-0003-1443-7900
atmdsmi@gmail.com
Kokand State University

Abstract: *Metadata is the connective tissue of scholarly communication, it determines whether a published article can be found, indexed, cited, and correctly attributed within the global research infrastructure. Yet the manual creation and verification of bibliographic metadata remains a slow, labour-intensive, and error-prone process, and this burden falls disproportionately on journals that publish outside the small set of languages and layout conventions for which existing extraction tools were designed. This paper examines the scholarly and economic case for automated, multilingual metadata extraction, using MetaExtract - a rule-based and named-entity-recognition hybrid system built for Uzbek (Latin and Cyrillic), Russian, and English academic journal articles - as a case study. Evaluated on a stratified 150-article gold-standard corpus, the system achieved an overall F1 score of approximately 0.965 across six core metadata fields, while processing a single document in an average of 2.62 seconds on modest, GPU-free hardware. We situate these results within the broader literature on metadata quality, discoverability, and the resource economics of extraction approaches, and argue that lightweight, format-aware hybrid architectures offer a more sustainable path to metadata automation for linguistically underserved journals than either purely manual workflows or resource-intensive large-model approaches.*

Keywords: *metadata extraction, scholarly communication, multilingual processing, digital libraries, bibliometrics, natural language processing, academic publishing economics*

1. Introduction

Academic articles are the principal currency of scientific exchange, and every year the volume of that currency grows, millions of new papers enter circulation annually, each carrying a bibliographic record - title, authors, affiliations, keywords, abstract, and identifiers such as the DOI - that allows it to be found, cited, and connected to the wider body of knowledge. This bibliographic record is metadata, and its accuracy and completeness are not incidental details of publishing but a direct determinant of a paper's visibility and impact. Recent large-scale evidence from institutional repositories confirms this link empirically, repository records with more complete metadata attract measurably higher user engagement than sparsely described ones (Rasuli et al., 2025). Metadata quality, in other words, is not a back-office concern; it is a variable that shapes whether research is read at all.

For most of the history of digital publishing, metadata has been produced by hand - typed into submission systems by authors, checked by editors, and re-keyed by indexing services. This manual pipeline is costly in aggregate, sensitive to human error, and difficult to scale as journal output grows, a concern raised explicitly in work on making large, messy document collections navigable through automated description (Skluzacek, 2022). Automated metadata extraction has therefore become an active research area at the intersection of information retrieval, computer vision, and computational linguistics. However, the great majority of extraction systems - including widely used open-source

tools such as GROBID and CERMINE - were designed and trained on English-language, Latin-script, digitally native documents from major international publishers (Tkaczyk, Szostek, Fedoryszak, Dendek, & Bolikowski, 2015). Journals that publish in other languages, use other scripts, or follow regional layout conventions are correspondingly underserved, even though the economic case for automation applies to them at least as strongly as it does to the journals such tools were built for.

This paper makes two connected arguments. First, drawing on the literature on metadata quality and discoverability, we argue that the scholarly cost of poor or missing metadata is substantial and falls hardest on journals and institutions with the least capacity to absorb it - typically those publishing in national or regional languages. Second, using MetaExtract, a hybrid rule-based and NER system built for Uzbek-, Russian-, and English-language journal articles, as an extended case study, we show that a modestly resourced, format-aware extraction architecture can achieve accuracy competitive with far more resource-intensive approaches while running on ordinary desktop hardware in a few seconds per document. Together, these findings support a broader claim - automated multilingual metadata extraction is not merely a technical convenience but an economically rational investment for national scientific publishing infrastructures, including in contexts such as Uzbekistan's, where journals publish in parallel scripts and languages and rarely have access to the metadata pipelines of large international publishers.

2. Metadata as Scholarly Infrastructure

Metadata performs several distinct but interlocking functions in the scientific communication system. At the most basic level, it enables discovery - search engines, library catalogues, and disciplinary repositories rely on title, author, keyword, and abstract fields to index and retrieve articles. When these fields are missing, inconsistent, or incorrectly parsed, an article effectively becomes harder to find even if the underlying research is sound. The empirical link between metadata completeness and downstream engagement documented by Rasuli et al. (2025) is one of the clearest demonstrations that this is not a theoretical concern but a measurable effect on how research is used.

Beyond discovery, metadata underpins the bibliometric and scientometric apparatus through which the scholarly community assesses influence and tracks the development of fields - citation counts, journal rankings, and co-authorship networks are all constructed from metadata records. Errors or omissions at the level of an individual article are therefore not self-contained; they can distort citation indices and institutional rankings at a much larger scale, an effect that is amplified for journals whose entire output shares a common non-standard layout or naming convention. This is a particularly acute issue for national scientific publishing systems such as Uzbekistan's, where academic journals appear in parallel Latin- and Cyrillic-script Uzbek editions as well as in Russian, and where a single extraction failure pattern can silently propagate across an entire journal run rather than affecting an isolated article.

Finally, metadata is a precondition for the FAIR principles - findable, accessible, interoperable, and reusable - that increasingly guide research-data and publication policy. None of these properties can be realised without structured, machine-actionable descriptive records. The case for treating metadata as scholarly infrastructure, rather than administrative overhead, follows directly from these functions, infrastructure failures are rarely visible at the point of failure, but they degrade the system's usefulness for everyone who depends on it.

3. The Hidden Cost of Manual Metadata Work

If metadata is infrastructure, its production is a labour cost that scales with publication volume. Manual metadata entry and verification requires a human - typically an editor, librarian, or indexing specialist - to read each article, identify the relevant fields, and transcribe them accurately, a process that is time-consuming even for a single, well-formatted document and considerably more so for large

or heterogeneous collections (Skluzacek, 2022). The problem compounds in exactly the settings where automation would be most valuable, retrospective digitisation of journal backlogs, migration to new publishing platforms such as Open Journal Systems, and ongoing processing of new issues across many journals simultaneously. In each of these scenarios, the number of documents to be processed can run into the hundreds of thousands, at which point even a modest per-document time cost accumulates into weeks or months of dedicated staff time.

Manual processes also carry a quality-control cost that is easy to underestimate. Human transcription is subject to fatigue-driven error, inconsistent application of formatting rules across staff members, and simple typographical mistakes - all of which are precisely the kinds of errors that downstream bibliometric and discovery systems cannot distinguish from genuine data. Correcting such errors after the fact, once they have propagated into citation databases or institutional repositories, is typically far more expensive than preventing them at the point of entry. The economic argument for automation therefore rests on two combined effects - the direct labour-hour savings from not manually transcribing metadata, and the avoided downstream cost of correcting errors that manual transcription would otherwise introduce. Well-designed automated pipelines, even those with only linear-time complexity in document length, have been shown in the digital-library literature to substantially outperform manual workflows on both dimensions (Herrera-Urtiaga, Torres-Cruz, Mendoza-Mollocondo, Paredes-Quispe, & Chambi-Mamani, 2023).

4. Automated Extraction: State of the Art and the Multilingual Gap

Research on automated metadata extraction has evolved through several methodological generations - early rule-based systems relying on regular expressions and layout heuristics; classical machine-learning classifiers such as support vector machines, k-nearest neighbours, and conditional random fields; and, most recently, deep neural architectures and large language models. Widely deployed open-source extractors such as GROBID and CERMINE report strong performance on the digitally native, English-language, standard-template documents for which they were built and evaluated (Tkaczyk et al., 2015), and heuristic, regular-expression-based approaches applied to well-defined scanned title pages have achieved field-level accuracy as high as 97% in similarly constrained settings (Choudhury, Wu, Ingram, & Fox, 2020).

This performance, however, degrades sharply outside the conditions for which such systems were tuned. Multimodal approaches that combine document layout analysis with visual features have been developed specifically because purely textual extraction proves unreliable once layout and typography vary across publishers, as demonstrated for German-language publications using a Mask R-CNN architecture trained on both natural and synthetic data (Boukhers, Beili, Hartmann, Goswami, & Zafar, 2021). Language- and script-specific challenges compound the problem further, research on metadata extraction from Persian-language theses and dissertations has shown that general-purpose tools trained on English morphology and structure transfer poorly to languages with substantially different naming conventions and word structures, motivating dedicated ensemble and two-stage classification approaches for that context (Rahnama, Hasheminejad, & Nasiri, 2020). No comparable body of tooling exists for Uzbek-language academic publishing, which additionally must accommodate parallel Latin- and Cyrillic-script editions alongside Russian-language articles within the same journals - a combination of multilingual and multi-script variation that falls outside the design assumptions of virtually all existing open-source extractors.

A further consideration, directly relevant to the economic argument developed in this paper, concerns the resource cost of different extraction approaches. Comparative work on automatically generating catalogue records from Turkish-language title pages found that a domain-tuned BERT-based model achieved higher accuracy (F1 = 0.666) than general-purpose vision-language models such as Gemini

2.5 Flash (F1 = 0.533), while also being markedly cheaper to run per processed sample in terms of both direct cost and estimated carbon footprint (Usta, 2026). The broader implication of that study - that specialised, narrowly scoped models can outperform large general-purpose models on both accuracy and resource economy for structured extraction tasks - motivates the architectural choices examined in the case study that follows.

5. MetaExtract: A Case Study in Multilingual Metadata Extraction

5.1 System Overview

MetaExtract is a Java-based desktop application built to extract structured metadata - title, author names, affiliation, keywords, abstract, and DOI - from academic journal articles published in Uzbek (Latin and Cyrillic script), Russian, and English. The system combines PDF text and layout extraction (via PDFBox and iText), optical character recognition for scanned pages (via Tess4J), and named-entity recognition (via Apache OpenNLP) within a rule-based architecture that additionally incorporates document-format signals - font size, weight, and relative position on the page - to disambiguate lines that are lexically similar but structurally distinct, for example a title continuation line versus an affiliation line. This format-aware 'veto' mechanism was developed specifically to address a systematic weakness identified during iterative testing, purely text-based heuristics for fields such as affiliation are prone to misclassifying position or title lines that happen to share vocabulary with genuine institutional names, an error mode consistent with observations in the wider literature that purely lexical or contextual features are insufficiently robust to layout diversity across publishers (Boukhers & Bouabdallah, 2022).

5.2 Evaluation Design

System accuracy was evaluated against a purpose-built gold-standard corpus of 150 journal articles, a sample size derived using Cochran's formula for an unknown population at a 95% confidence level and an 8% margin of error, and consistent with sample sizes used in comparable digital-library evaluation studies (Choudhury et al., 2020). The corpus was drawn from the open archives of 15 Uzbek scientific journals across 57 journal-issue combinations, spanning technical, natural-science, social-science, and pedagogical disciplines, and was stratified by language (82 Uzbek, 37 Russian, 31 English articles), script (45 Latin- and 37 Cyrillic-script Uzbek articles), and title-author block layout (77 title-first, 73 author-first articles) to ensure that evaluation results reflect genuine publishing diversity rather than a single dominant layout pattern. Gold-standard values for each field were manually verified against both the article text and the corresponding Open Journal Systems metadata record prior to running the automated system, and system output was compared against this reference using standard precision, recall, and F1 metrics at the level of individual metadata fields.

5.3 Accuracy Results

Across all six evaluated fields and all three languages, MetaExtract achieved an overall F1 score of approximately 0.965 - a result competitive with figures reported for comparable heuristic and hybrid systems evaluated on more linguistically homogeneous corpora (Choudhury et al., 2020; Choudhury, Jayanetti, Wu, Ingram, & Fox, 2021). Field-level results are summarised in Table 1.

Table 1.

Field-level precision, recall, and F1 scores on the 150-article gold-standard corpus.

Field	Precision	Recall	F1
Title	0.987	0.987	0.987
Author names	1.000	0.990	0.995
Affiliation	0.974	0.921	0.947
Abstract	0.960	0.953	0.957

Field	Precision	Recall	F1
Keywords	0.960	0.927	0.943
DOI	0.967	0.967	0.967

Title and author-name detection were the most reliable fields, reflecting the effectiveness of combining format-based cues with named-entity recognition for exactly the elements those two signal sources complement each other on. Keywords produced the lowest field-level score, consistent with the greater formatting variability of that field across journals - commas, semicolons, and separate lines are all used inconsistently as delimiters, and the label preceding the field itself varies in wording. Performance also varied moderately by language - Russian-language articles achieved the highest average F1 (≈ 0.990), attributable to the comparatively stable patronymic-based author-naming convention in Russian academic writing, while Uzbek and English articles - affected respectively by possessive-suffixed surname variants and by larger, more heterogeneous international author lists - scored slightly lower, though still above 0.94 overall. Error analysis indicated that the majority of remaining false positives stemmed from text-only affiliation heuristics misclassifying position or title lines, an error class substantially reduced but not eliminated by the format-control mechanism, while most false negatives involved atypical or abbreviated author-name forms not anticipated by the original naming-pattern rules.

5.4 Processing Speed and Computational Economy

Beyond accuracy, the practical and economic value of an extraction system depends on how quickly and cheaply it runs. Using purpose-built timing instrumentation, MetaExtract processed a single document - from PDF load through full field extraction and export - in an average of 2.62 seconds (median 2.37 seconds) on a modest, GPU-free machine (Intel Core i5-2400, 8 GB RAM). The full 150-article corpus was processed sequentially in a single instance in 392.9 seconds, or approximately 6.55 minutes. Regression analysis showed processing time to be almost independent of document length (Pearson $r = 0.09$), the dominant cost is a fixed per-document overhead of roughly 2.45 seconds for model and library initialisation, with each additional page adding only about 28 milliseconds, a pattern also reported for deep-learning-based extraction pipelines, where model-loading time frequently exceeds the time required to process an individual document once loaded (Safder et al., 2020).

This combination of accuracy and speed has a direct economic reading. A processing time measured in seconds per document, rather than the minutes required for careful manual transcription, implies that a single ordinary desktop machine could process a national journal's entire annual output, or a substantial retrospective digitisation backlog, well within a normal working day, and that the marginal cost of processing an additional document is close to zero once the system is deployed. Because MetaExtract's architecture applies its heaviest computational component - named-entity recognition - only where it is strictly necessary, reusing font and layout information already produced as a by-product of PDF parsing rather than invoking a separate visual-processing pipeline, it avoids the GPU and infrastructure requirements associated with deep-learning and vision-language extraction approaches (Usta, 2026). It is, moreover, naturally suited to further scaling, because each document is processed independently of every other, the underlying workload is embarrassingly parallel, and larger backlogs can in principle be distributed across multiple ordinary machines or a modest computing cluster with speed-up roughly proportional to the number of parallel instances, following the general pattern documented for serverless, distributed metadata-extraction frameworks operating at far larger scale (Skluzacek, Wong, Li, Chard, Chard, & Foster, 2021).

6. Discussion: Quantifying the Economic Benefit

Taken together, the accuracy and performance results above support a straightforward economic argument. First, on the labour side, a system that extracts six metadata fields at 0.965 average F1 in roughly 2.6 seconds per document removes the great majority of manual transcription work from journal editors, indexers, and librarians, leaving only a residual verification task focused on the smaller number of fields - keywords and, to a lesser extent, affiliation - where accuracy is somewhat lower and human review remains valuable. Because verification of a machine-generated draft is considerably faster than transcription from scratch, the realistic labour saving is larger than the raw F1 figures alone would suggest.

Second, on the infrastructure side, the comparison with resource-intensive approaches is instructive. Where deep neural and vision-language models can achieve strong accuracy on well-resourced, high-volume languages, they typically require specialised hardware and carry meaningfully higher per-sample cost and environmental footprint than smaller, task-specific models (Usta, 2026). For national and regional scientific publishing systems, which typically operate with limited technical budgets and cannot justify investment in GPU infrastructure for a comparatively modest annual publication volume, a lightweight, rule-and-NER hybrid architecture that runs on ordinary office hardware represents a materially more sustainable investment. This is precisely the gap in which languages such as Uzbek, published in parallel scripts alongside Russian, currently sit - underserved by English-centric open-source tools such as GROBID and CERMINE, and unable to economically justify the infrastructure that large general-purpose models require.

Third, the scalability characteristics documented in Section 5.4, near-zero marginal cost per additional document, and an embarrassingly parallel workload, mean that the economic benefit compounds with journal output rather than diminishing. A system that is merely convenient for a single journal issue becomes economically decisive once applied across a multi-year retrospective digitisation project or a country-wide network of journals migrating to a shared publishing platform such as Open Journal Systems. In such settings, the avoided cost is not measured in minutes per article but in staff-months or staff-years across an entire national publishing system, alongside the harder-to-quantify but empirically supported benefit of improved discoverability and citation visibility that follows from more complete metadata (Rasuli et al., 2025).

7. Limitations and Future Work

This case study has several limitations that bound the strength of its conclusions. The gold-standard corpus, while stratified by language, script, and layout, was drawn exclusively from Uzbek journal archives, and the reported results have not yet been benchmarked directly against GROBID or CERMINE on the same corpus; such a head-to-head comparison would allow the relative advantage of a language-specific hybrid architecture to be quantified rather than inferred from separate studies conducted on different corpora. Likewise, no direct comparison has been made between MetaExtract and large language model-based extraction on the same documents, which would allow the economic argument advanced here to be tested with matched accuracy figures rather than by analogy to comparable studies in other languages (Usta, 2026). Finally, the parallel-processing potential of the architecture, while theoretically supported by its document-independent processing model, has not yet been tested in a genuinely distributed or multi-node environment, and remains an open direction for subsequent work, alongside a formal cost-accounting study translating the time savings documented here into concrete labour-cost estimates for a specific journal or library system.

8. Conclusion

Metadata is not a peripheral detail of academic publishing but a structural determinant of whether research can be found, cited, and integrated into the wider scientific record, and the cost of producing it manually is a real and scalable burden that falls hardest on journals working outside the linguistic

and technical mainstream of international publishing. This paper has argued, and illustrated through the MetaExtract case study, that a modestly resourced, format-aware hybrid extraction architecture can close much of this gap: achieving accuracy competitive with far more resource-intensive systems, running in a few seconds per document on ordinary hardware, and scaling economically to the volumes involved in national journal digitisation and platform migration. For multilingual and script-diverse publishing environments such as Uzbekistan's, where existing English-centric open-source tools offer limited direct support, this combination of accuracy, speed, and low infrastructure cost represents not merely a technical convenience but an economically sound investment in the country's scientific communication infrastructure.

References

[1] Boukhers, Z., & Bouabdallah, A. (2022). Vision and natural language for metadata extraction from scientific PDF documents: A multimodal approach. In Proceedings of the 2022 ACM/IEEE Joint Conference on Digital Libraries (JCDL) (pp. 1–5). Cologne, Germany.

[2] Boukhers, Z., Beili, N., Hartmann, T., Goswami, P., & Zafar, M. A. (2021). MexPub: Deep transfer learning for metadata extraction from German publications. In Proceedings of the 2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL) (pp. 250–253). Champaign, IL, USA. https://doi.org/10.1109/JCDL52503.2021.00076

[3] Choudhury, M. H., Wu, J., Ingram, W. A., & Fox, E. A. (2020). A heuristic baseline method for metadata extraction from scanned electronic theses and dissertations. In Proceedings of the ACM/IEEE Joint Conference on Digital Libraries 2020 (JCDL '20) (pp. 515–516). Association for Computing Machinery. https://doi.org/10.1145/3383583.3398590

[4] Choudhury, M. H., Jayanetti, H. R., Wu, J., Ingram, W. A., & Fox, E. A. (2021). Automatic metadata extraction incorporating visual features from scanned electronic theses and dissertations. In Proceedings of the 2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL) (pp. 230–233). Champaign, IL, USA. https://doi.org/10.1109/JCDL52503.2021.00066

[5] Herrera-Urtiaga, A. P., Torres-Cruz, F., Mendoza-Mollocondo, C. I., Paredes-Quispe, J. R., & Chambi-Mamani, E. W. (2023). Automatic extraction of metadata based on natural language processing for research documents in institutional repositories. In Proceedings of the 8th Brazilian Technology Symposium (BTSym '22). Smart Innovation, Systems and Technologies.

[6] Rahnama, M., Hasheminejad, S. M. H., & Nasiri, J. A. (2020). Automatic metadata extraction from Iranian theses and dissertations. In Proceedings of the 2020 6th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS) (pp. 1–5). Mashhad, Iran. https://doi.org/10.1109/ICSPIS51611.2020.9349570

[8] Rasuli, B., Boock, M., Schöpfel, J., et al. (2025). The link between dissertation metadata completeness and user engagement in an institutional repository. Scientometrics, 130, 2875–2899. https://doi.org/10.1007/s11192-025-05331-0

[9] Safder, I., Hassan, S.-U., Visvizi, A., Noraset, T., Nawaz, R., & Tuarob, S. (2020). Deep learning-based extraction of algorithmic metadata in full-text scholarly documents. Information Processing & Management, 57(6), 102269.

[10] Skluzacek, T. J. (2022). Automated metadata extraction can make data swamps more navigable (Doctoral dissertation). The University of Chicago.

Skluzacek, T. J., Wong, R., Li, Z., Chard, R., Chard, K., & Foster, I. (2021). A serverless framework for distributed bulk metadata extraction. In Proceedings of the 30th International Symposium on High-Performance Parallel and Distributed Computing (pp. 7–18).

Tkaczyk, D., Szostek, P., Fedoryszak, M., Dendek, P. J., & Bolikowski, L. (2015). CERMINE: Automatic extraction of structured metadata from scientific literature. International Journal on Document Analysis and Recognition, 18(4), 317–335.
Usta, G. (2026). Automating MARC records from Turkish title pages: A performance and efficiency comparison of fine-tuned BERT vs. VLMs. Journal of Library Metadata, 1–19. https://doi.org/10.1080/19386389.2026.2627732