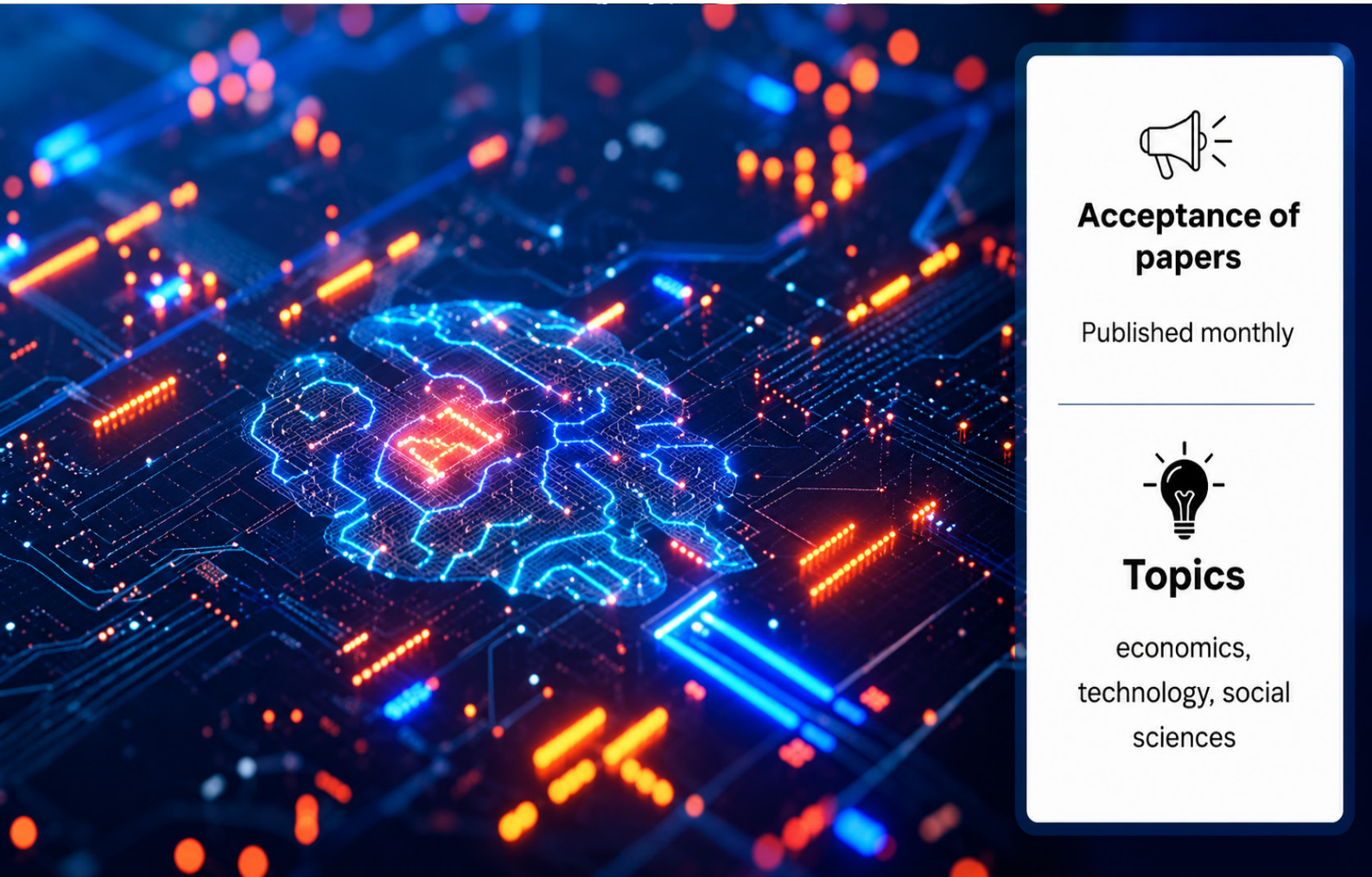
TASHKENT STATE UNIVERSITY OF ECONOMICS

International journal of topics

INNOVATION SCIENCE AND TECHNOLOGY

ISSUE 9, PART I / SEPTEMBER 2026

Acceptance of papers



Acceptance of papers

Published monthly

Topics

economics, technology, social sciences

INNOVATION SCIENCE AND TECHNOLOGY

Google Scholar

Digital Object Identifier

RePEc RESEARCH PAPERS IN ECONOMICS

ISSN 3060-5229

Electronic publication, Issue 9.
Approved for publication on September, 2026.

THE SCIENTIFIC-POPULAR ELECTRONIC JOURNAL **"INNOVATION SCIENCE AND TECHNOLOGY"** HAS BEEN REGISTERED UNDER THE NUMBER **C-5669633** BY THE AGENCY FOR INFORMATION AND MASS COMMUNICATIONS (AOKA) OF THE REPUBLIC OF UZBEKISTAN, EFFECTIVE FROM OCTOBER 9, 2024.

CONTACTS

Phone: **+998 50 737 87 88**

Website: https://ist-journal.uz

Email: innovationist2025@gmail.com

The scientific electronic journal “Innovation Science and Technology” has been included in the list of scientific publications recommended for the publication of main scientific results of dissertations for the award of PhD and DSc degrees in economics and technical sciences, in accordance with the Resolution No. 370 of the Presidium of the Higher Attestation Commission of the Republic of Uzbekistan, dated May 8, 2025.

Editorial board:

CONTENTS

METAEXTRACT: A HYBRID METADATA EXTRACTION SYSTEM COMBINING RULE-BASED AND STATISTICAL APPROACHES

Rashid Muhtorovich Turgunbaev
Associate Professor, Department of Digital Technologies and Artificial Intelligence
Kokand State University
ORCID: 0000-0003-1443-7900
Email: atmdsmi@gmail.com

Abstract. The issue of automatic extraction of structural metadata from documents is a widely studied topic in the field of document analysis, in which two main paradigms — universal models based on deep learning and rule-based systems — compete. Universal models have high generalization capabilities, but they require large amounts of annotated training data, which is a serious limitation for low-resource languages and domains. In this paper, a hybrid architecture combining a rule-based geometric-textual approach with a statistical named entity recognition (NER) model is presented through the example of extracting bibliographic metadata from scientific journal articles. The proposed system is based on the principle of creating rules from a single sample document and applying them to other documents sharing the same template, which makes it practically applicable even without large amounts of training data. The paper discusses design principles, the justification of architectural decisions, implementation details, and the role of this approach in the broader field of document extraction.

Keywords: document metadata extraction; rule-based systems; hybrid architecture; named entity recognition (NER); low-resource languages; software design.

Annotatsiya. Hujjatlardan strukturaviy metama'lumotlarni avtomatik ekstraksiya qilish masalasi hujjat tahlili sohasida keng o'rganilgan mavzu bo'lib, unda ikkita asosiy paradigma — chuqur o'rganishga asoslangan universal modellar va qoidalarga asoslangan tizimlar — raqobatlashadi. Universal modellar yuqori umumlashtirish qobiliyatiga ega, biroq ular yirik hajmdagi annotatsiyalangan o'quv ma'lumotlarini talab qiladi, bu esa kam resursli tillar va domenlar uchun jiddiy cheklov hisoblanadi. Ushbu maqolada ilmiy jurnal maqolalaridan bibliografik metama'lumotlarni ekstraksiya qilish misolida qoidalarga asoslangan geometrik-matnli yondashuv bilan statistik nomlangan mavjudotlarni aniqlash (NER) modelini birlashtirgan gibrid arxitektura taqdim etiladi. Taklif etilgan tizim bitta namunaviy hujjat asosida qoidalar yaratish va ularni bir xil andozadagi qolgan hujjatlarga qo'llash tamoyiliga asoslanadi, bu esa uni katta hajmdagi o'quv ma'lumotlarisiz ham amaliy jihatdan qo'llanuvchan qiladi. Maqolada dizayn tamoyillari, arxitekturaviy qarorlarning asoslanishi, amalga oshirish tafsilotlari hamda yondashuvning umumiy hujjat-ekstraksiya sohasidagi o'rni muhokama qilinadi.

Kalit so'zlar: hujjatdan metama'lumot ekstraksiya qilish; qoidalarga asoslangan tizimlar; gibrid arxitektura; nomlangan mavjudotlarni aniqlash (NER); kam resursli tillar; dasturiy dizayn.

Аннотация. Вопрос автоматического извлечения структурных метаданных из документов является широко изучаемой темой в области анализа документов, в которой конкурируют две основные парадигмы — универсальные модели, основанные на глубоком обучении, и системы, основанные на правилах. Универсальные модели обладают высокими обобщающими возможностями, но требуют больших объемов аннотированных обучающих данных, что является серьезным ограничением для языков и предметных областей с ограниченными ресурсами. В данной статье представлена гибридная архитектура, сочетающая геометрико-текстовый подход, основанный на правилах, со статистической моделью распознавания именованных сущностей (NER) на примере извлечения библиографических метаданных из статей научных журналов. Предложенная система основана на принципе создания правил на основе одного образца документа и применения их к другим документам того же шаблона, что делает ее практически применимой даже без большого объема обучающих данных. В статье обсуждаются принципы проектирования, обоснование архитектурных решений, детали реализации и роль подхода в общей области извлечения данных из документов.

Ключевые слова: извлечение метаданных из документов; системы, основанные на правилах; гибридная архитектура; распознавание именованных сущностей (NER); языки с ограниченными ресурсами; проектирование программного обеспечения.

INTRODUCTION

Automatically extracting structural metadata from various standardized documents (scientific papers, invoices, contracts, technical documents) is one of the key areas in the field of information extraction. In this domain, **an optimal balance** is typically sought between high accuracy and low resource consumption. Artificial intelligence capabilities were also utilized during the editing process of this text. Although universal extractors based on deep neural networks possess high generalization capabilities, they require large amounts of annotated training corpora. Since such ready-made data is not sufficiently available for many languages and narrow domains, the need arises to employ alternative approaches.

Rule-based systems, on the other hand, ensure high accuracy across document collections with a specific template and do not require large amounts of training data; however, due to their reliance on a single template, they differ from universal models in terms of flexibility. This paper presents a hybrid architecture combining both approaches—using the practical task of extracting bibliographic metadata from scientific journal articles as an example—and discusses its design principles.

LITERATURE REVIEW

Automatically extracting bibliographic metadata—such as title, authors, affiliation, abstract, and keywords—from scientific journal articles is essential for populating digital libraries and scientific databases (such as Scopus, Web of Science, and Google Scholar). Since manual data entry is time-consuming, prone to errors, and resource-intensive, various software tools have been developed over the past decades to automate this task. Artificial intelligence capabilities were also utilized during the editing process of this text.

Lopez and Romary presented the GROBID system, which extracts metadata and bibliographic references from PDF documents using sequence labeling models based on conditional random fields and generates TEI-structured outputs. Tkaczyk et al. described the CERMINE system, which employs a modular architecture and machine learning algorithms that account for both text content and geometric features to extract expanded bibliographic fields like journal name, volume, and issue. Additionally, systems such as ParsCit, Neural ParsCit, Science Parse, and PdfAct utilize varying combinations of rule-based and statistical methods.

Comparative studies evaluating these systems on diverse document collections highlight their respective strengths. On the DocBank benchmark, GROBID achieved top performance in metadata and reference extraction, followed by CERMINE and Science Parse, while noting common challenges across all tools regarding tables, lists, and footnotes. In an evaluation of Indonesian scientific publications, AnyStyle and CERMINE delivered strong overall results, though performance varied by domain. These findings demonstrate that effectiveness depends heavily on document layout, language, and journal formatting.

Many researchers emphasize that existing tools face limitations when processing scanned documents or non-Latin scripts. This affects their direct applicability to regional scientific journals published in Cyrillic and Latin scripts across Uzbek, Russian, and English. Consequently, interest has grown in hybrid—neuro-symbolic—approaches, particularly for low-resource languages. In these architectures, rule-based components simplify initial segmentation or label complexity, while statistical or deep learning models perform high-precision final classification. For instance, Duc et al. proposed a two-stage hybrid system for Vietnamese that reduces label complexity using rules before fine-tuning pre-trained language models, achieving notable gains over purely neural baselines.

Modern Approaches to Metadata Extraction

Modern extraction architectures generally rely on two primary paradigms: machine learning/deep neural network models designed for universal extraction (e.g., GROBID, CERMINE) and rule-based systems optimized for structured, template-specific documents. These approaches are complementary and frequently integrated into hybrid systems, where geometric and marker-based rules offer rapid, accurate preliminary extractions, and statistical or neural models provide secondary verification. Beyond academic literature, rule-based extraction is widely applied to standardized commercial documents, such as invoices, confirming its practical reliability. Standard metrics—precision, recall, and F1-score—are systematically used to evaluate and compare metadata extraction performance.

RESEARCH METHODOLOGY

MetaExtract is developed in the Java programming language as a desktop application with a Swing graphical user interface. The system utilizes several key libraries: Apache PDFBox for reading the text layer and page structure (layout, font, coordinates) of PDF documents; iText for supplementary PDF processing; Tess4J for optical character recognition (OCR) to convert scanned pages into text; Apache OpenNLP for statistical named entity recognition; and the Jackson library for serializing and deserializing rule files and configurations in JSON format.

ANALYSIS AND RESULTS

The core concept of the system involves integrating two complementary components:

Rule-based component: A «format-check/veto» mechanism based on font and layout profiles learned for each journal. This component evaluates text-matching rules in conjunction with formatting properties (font size, weight, position), thereby reducing misclassification errors common in purely text-based heuristics (such as incorrectly identifying a job title or header line as an affiliation). Artificial intelligence capabilities were also utilized during the editing process of this text.

Statistical component: A named entity recognition (NER) model trained using the Apache OpenNLP library on a specialized corpus in Uzbek (Latin and Cyrillic) and Russian, which enables the detection of author names even in cases with missing space delimiters or atypical formatting patterns.

To enable joint operation, the system supports three execution modes: rule-based only, statistical only, and combined. In the combined mode, the rule-based component acts as a filter that validates or rejects (vetoes) the predictions of the statistical model, thereby improving overall accuracy. To semi-automate the creation of journal-specific rules instead of manual coding, a companion GUI tool, PDFTextLabeler, was developed. This tool allows users to visually annotate PDF document elements and automatically generates a JSON rule file containing font and layout profiles.

To objectively evaluate system performance, a stratified gold standard corpus of 150 scientific articles was constructed across three dimensions: (1) language (Uzbek Latin, Uzbek Cyrillic, Russian, English), (2) name format patterns, and (3) layout types. This stratification ensures proportional representation across different languages and document formats, enhancing the generalizability of evaluation results. During corpus construction, multi-stage verification was conducted to eliminate duplicate articles, annotation gaps, and identifier conflicts.

A comprehensive gold-standard evaluation framework was implemented in Java to compare system-extracted metadata against gold-standard references. Alignment is performed based on normalized titles, with fuzzy matching applied when necessary. Field-specific (title, authors, affiliation, etc.) and language-specific F1-scores are computed independently. Additionally, the integrated GoldEvaluator tool exports these results to CSV format.

To assess operational efficiency, execution time instrumentation was incorporated at both document and processing-stage levels, recording real-time latency data in CSV format for performance analysis.

Proposed Hybrid Architecture

The proposed hybrid architecture consists of two interconnected applications—a rule creation tool and an extraction tool—linked via a shared JSON configuration schema. The core principle relies on the structural uniformity of documents from a single source (e.g., a specific journal). Users annotate metadata fields on a single sample document using geometric rectangle and line modes, enabling the system to automatically generate a corresponding rule set (Table 1). The resulting rules are subsequently applied across all other documents sharing the same layout template.

Table 1

Core Technologies Used in the System

Layer	Technology	Function / Role
Document Ingestion	Apache PDFBox (Primary), iText (Fallback)	Extracting text and geometric information (x, y, font) from PDF documents.
Scanned Documents	Tess4J (Tesseract OCR)	Converting image-only pages without text layers into searchable text.
Metadata Extraction	Apache OpenNLP (langdetect, NameFinderME)	Providing supplementary statistical (NER) signals for author name verification. (Note: Artificial intelligence capabilities were also utilized during the editing process of this text).
Configuration & Storage	Jackson (ObjectMapper)	Serializing and deserializing ExtractionConfig objects to/from JSON format.
User Interface	Java Swing	Providing the graphical user interface for PDFTextLabeler and MetaExtract.
Development & Build	Apache Ant, NetBeans	Packaging the project into a unified executable JAR file and managing dependencies.

To enhance the reliability of rule definitions, the system integrates two types of extraction signals: geometric-positional signals (coordinates, font size, alignment) and textual-marker signals (keyword anchors). Additionally, for complex tasks such as author name extraction, a statistical Named Entity Recognition (NER) model acts as a secondary validation layer. Artificial intelligence capabilities were also utilized during the editing process of this text. This hybrid setup reduces dependency on positional parameters alone by incorporating linguistic features. Combining these three signals (geometric, marker, and NER) mitigates individual failure modes: geometric rules alone may fail across layout variations, while relying solely on marker keywords can introduce ambiguity when keywords recur within the document text (Figure 1).

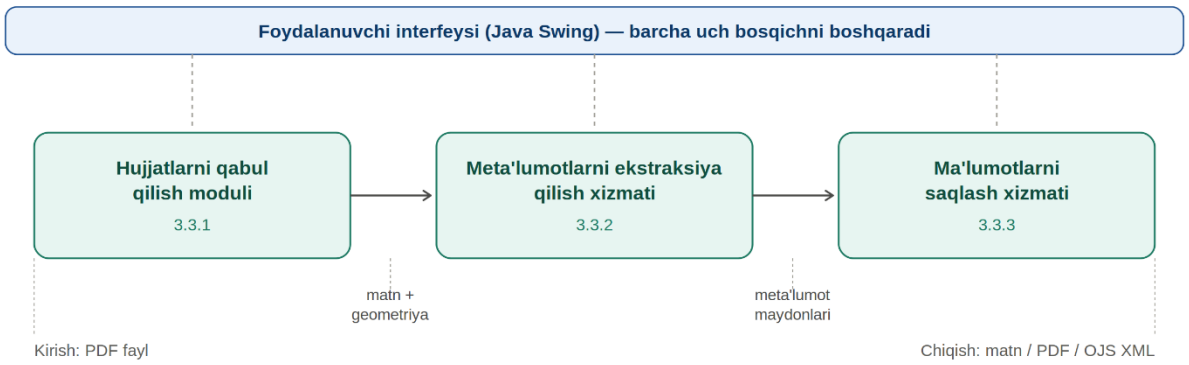


Figure 1. Data flow between modules in the hybrid architecture: from raw document to text/geometry, through metadata fields based on geometric+marker+NER signals, and ultimately to output files.

Implementation Details

From an implementation perspective, the system utilizes a primary standard library and a secondary fallback library at the document ingestion layer to extract text and geometric information. Artificial intelligence capabilities were also utilized during the editing process of this text. This dual-library strategy enables the system to process a broader spectrum of documents from diverse sources, as certain files—particularly those generated by older publishing software or non-standard text encodings—may yield incomplete or inaccurate parsing results when processed solely by the primary engine. For documents lacking an embedded text layer (such as scanned pages), the system automatically routes the workflow through an Optical Character Recognition (OCR) pipeline (Figure 2).

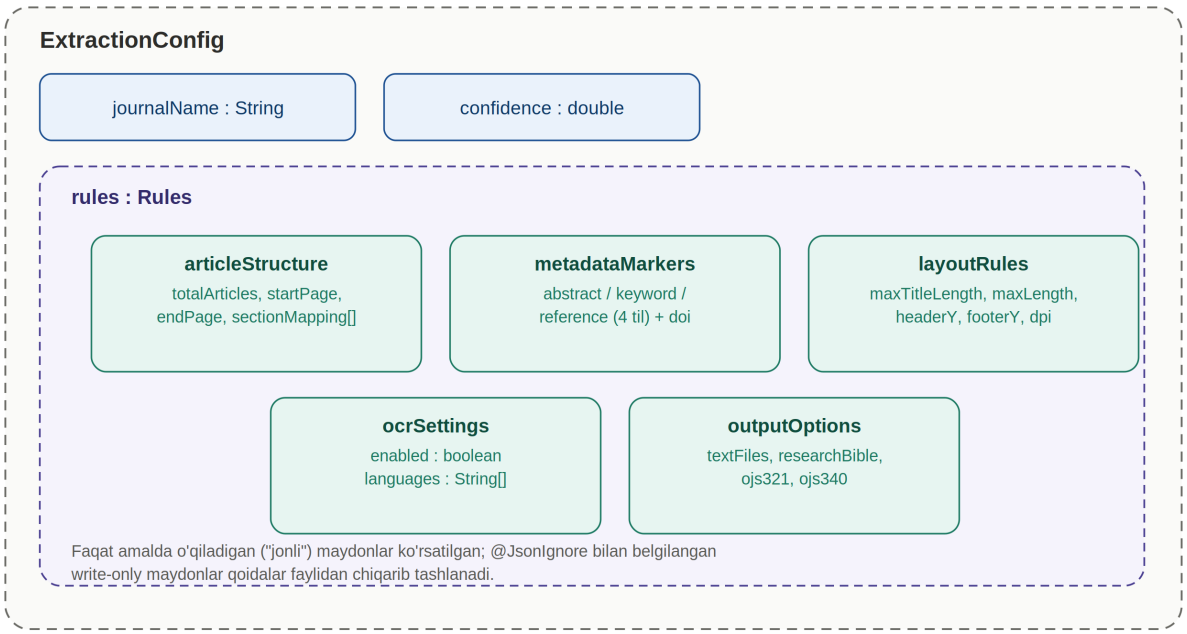


Figure 2. ExtractionConfig schema architecture: top-level metadata properties (journalName, confidence) and five sub-schemas (articleStructure, metadataMarkers, layoutRules, ocrSettings, outputOptions) nested within the rules

block.

English Translation & Academic Text Refinement

The configuration schema is implemented as a collection of object-oriented programming classes and serialized into JSON format using standard serialization libraries. An architectural advantage of this design is backward compatibility: adding new fields to the schema allows older configuration files to load seamlessly with default values, ensuring long-term software maintainability and scalability.

Artificial intelligence capabilities were also utilized during the editing process of this text. A practical challenge identified during implementation was the inconsistent recognition of typographical variants (e.g., multiple Unicode representations of apostrophes and modifiers in Uzbek). This highlights the necessity of applying a unified text normalization function across all matching nodes—a key design consideration for text processing systems handling low-resource languages.

The primary advantage of the proposed hybrid approach is its ability to resolve the traditional trade-off between high extraction accuracy and low resource expenditure. The system operates effectively without requiring large annotated training datasets, as primary extraction relies on geometric and marker-based signals while the statistical model serves strictly as a secondary validation layer. This creates a practical solution particularly suitable for low-resource languages and small-scale document archives where the data required to train or fine-tune universal models is often unavailable. Artificial intelligence capabilities were also utilized during the editing process of this text.

However, the approach carries an inherent constraint: because rules are bound to specific structural templates, the system cannot automatically adapt to documents with significantly different visual layouts, necessitating a separate rule generation process for each distinct template. This trade-off reflects the classic balance between the broad generalization capabilities of universal models and the high precision of rule-based systems.

CONCLUSION AND RECOMMENDATIONS

This paper presented a hybrid metadata extraction architecture that combines a rule-based geometric-textual approach with a statistical NER signal, illustrating its design principles through the practical task of extracting bibliographic metadata from scientific journal articles. Artificial intelligence capabilities were also utilized during the editing process of this text. It demonstrated that a rule-based approach founded on the «annotate once, apply many» principle—when augmented with a statistical model—delivers practically reliable results without requiring large amounts of annotated training data. Such an architecture serves as a viable, resource-efficient alternative to universal models for low-resource languages and narrow domains.

REFERENCES

1. Romary, L., & Lopez, P. (2015). GROBID — Information extraction from scientific publications. ERCIM News, 2015(100).
2. Tkaczyk, D., Szostek, P., Fedoryszak, M., Dendek, P. J., & Bolikowski, Ł. (2015). CERMINE: Automatic extraction of structured metadata from scientific literature. International Journal on Document Analysis and Recognition (IJDAR), 18(4), 317–335. https://doi.org/10.1007/s10032-015-0249-8
3. Duc, D. M., Truong, Q. X., Hong, V. T., Anh, L. H., Tra, M. T. M., Van Thuy, N., ... & Van, V. N. (2026). A hybrid method for low-resource named entity recognition. arXiv preprint arXiv:2605.04489.
4. Ahmed, M. W., & Afzal, M. T. (2020). FLAG-PDFe: Features oriented metadata extraction framework for scientific publications. IEEE Access, 8, 99458–99469. https://doi.org/10.1109/ACCESS.2020.2996112
5. Skluzacek, T. J., Chard, K., & Foster, I. (2022). Automated metadata extraction: Challenges and opportunities. In 2022 IEEE 18th International Conference on e-Science (e-Science) (pp. 495–500). IEEE. https://doi.org/10.1109/eScience56278.2022.00078
6. Ubewikkrama, P. T. Automatic invoice data identification with relations [Doctoral dissertation].
7. Skluzacek, T. J., Chen, M., Hsu, E., Chard, K., & Foster, I. (2022). Models and metrics for mining meaningful metadata. In International Conference on Computational Science (pp. 417–430). Springer, Cham. https://doi.org/10.1007/978-3-031-08757-8_35
8. Raval, P., & Bhaidasna, H. (2025). Metadata extraction from scholarly document using deep learning. In 2025 3rd International Conference on Inventive Computing and Informatics (ICICI) (pp. 1–5). IEEE.
9. Boukhers, Z., Beili, N., Hartmann, T., Goswami, P., & Zafar, M. A. (2021). MExPub: Deep transfer learning for metadata extraction from German publications. In 2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL) (pp. 250–253). IEEE. https://doi.org/10.1109/JCDL52503.2021.00038
10. Skluzacek, T. J. Automated metadata extraction can make data swamps more navigable [Doctoral dissertation, The University of Chicago].

Proofreader: Zokir ALIBEKOV

Layout and Designer: Oloviddin Sobir ugli

2026. № 9

To publish articles in this journal, you may submit articles, advertisements, stories, and other creative materials through the following links. Materials and advertisements are published on a paid basis.

You may subscribe to the journal at any time using the following details. Once subscribed, please send a screenshot or photo of your payment confirmation to our Telegram page @iqtisodiyot_77. Based on this, we will send the latest issue of the journal to your address each month.

"The journal "INNOVATION SCIENCE AND TECHNOLOGY" has been registered by the Agency for Information and Mass Communications under the Administration of the President of the Republic of Uzbekistan from 09.10.2024 under the registration number №390637. License number: C-5669633. PNFL: 30407832680027

Our address: Tashkent city, Yunusobod district, 19th block, House 17.

TASHKENT STATE UNIVERSITY OF ECONOMICS

Acceptance of articles

Published every monthly

Directions

Social, economic, political, technological, scientific

Scopus || Scientific electronic journal specializing in Scopus

CERTIFICATE NUMBER: №390637

ORDER NUMBER ACCORDING TO THE LICENSE REGISTER: C-5669633

CONTACT:

Contact us
+998 50 737 87 88

Telegram channel
t.me/scopus_IST2100

Journal official website
https://ist-journal.uz/index.php/IST