GDAŃSK UNIVERSITY OF TECHNOLOGY

TASK Quarterly 30 (1) 2026

Digital Paper: How PDF Stalled the Evolution of Documents

Magdalena Godlewska *

Gdańsk University of Technology, Faculty of Electronics, Telecommunications and Informatics, Dept. of Intelligent Interactive Systems, Gabriela Narutowicza 11/12, 80-233 Gdańsk, Poland

* maggodle@pg.edu.pl

https://doi.org/10.34808/tq2026/30.1/a

Abstract

Despite its limited support for semantic representation, the PDF format remains the dominant medium for document publication due to its visual fidelity, portability, and widespread acceptance. As a result, contemporary document-processing workflows often treat PDFs as digital paper rather than structured data, creating a persistent gap between human-oriented presentation and machine-based processing requirements.

PDF was designed primarily to preserve visual appearance, not document semantics or logical structure. Consequently, information extraction from PDF files is inherently complex and error-prone, as it requires reconstructing content hierarchy, relationships, and reading order from low-level layout cues. This limitation significantly constrains automated analysis, information retrieval, and AI-driven processing, particularly in domains such as law and public administration.

This paper presents a PDF processing approach developed within the CAISE project that transforms arbitrary text-based PDF documents into a structured JSON representation. The solution preserves layout-related metadata, spatial localisation of elements, and embedded resources such as images and tables, while addressing character encoding corruption commonly found in real-world PDFs. The resulting structured data provide a foundation for reconstructing logical document structure and enabling downstream semantic processing.

The proposed approach illustrates how presentation-oriented documents can be systematically transformed into machine-processable representations, mitigating the long-standing limitations of PDF-based workflows and supporting advanced document analysis and knowledge extraction.

Keywords:

PDF processing; document semantics; logical document structure; information extraction

1. Introduction

The PDF format [1] is one of the most widely used document formats for digital content distribution across public administration, industry, academia, and everyday communication. Its long-standing presence and broad software support have made it a de facto standard for document publication, particularly in contexts where visual consistency and platform independence are considered essential.

For non-technical users, PDF is especially attractive due to its stable presentation, predictable layout, and ease of sharing across different devices and operating systems. Documents published in PDF format appear identical regardless of the viewing environment, reinforcing the perception of reliability and professionalism. As a result, PDF is often regarded as a safe and final form of publication, suitable for long-term storage and official communication [2].

However, what is typically overlooked by non-technical authors is that publishing a document in PDF format fundamentally alters the nature of the document itself. Internally, PDF does not explicitly preserve the logical or semantic structure of document content. Instead, it represents documents as collections of precisely positioned visual elements—text fragments, glyphs, and graphical objects—optimised solely for accurate visual rendering. The semantic relationships between elements, such as content hierarchy, functional roles, or even the intended reading order of words, are not explicitly encoded and may be partially or entirely lost. From a machine-processing perspective, a PDF document resembles a visual artefact rather than a structured information resource.

In addition to misconceptions about structure, PDF is often perceived as a secure format, particularly in comparison to editable document types. This perception is misleading. The PDF specification allows for the inclusion of executable content, including scripts written in JavaScript, interactive forms, and embedded multimedia. While such features are rarely used in standard publishing workflows, their existence contradicts the widespread belief that PDF is inherently passive or immune to security-related concerns.

These issues are not merely theoretical. A prominent example can be observed in the Polish legislative system, where legal acts have historically been published primarily in PDF format without sufficient consideration of the implications for accessibility and machine readability [3]. Legislators and publishing authorities did not initially perceive PDF-based publication as problematic, given its visual clarity and formal appearance. Over time, however, this practice has resulted in severely limited access to legal information, particularly for citizens seeking to navigate, analyse, or cross-reference complex legal texts.

From the perspective of citizens, navigating hundreds or thousands of legal documents published as PDF files is a demanding and error-prone task. The lack of explicit structure, identifiers, and semantic links significantly hinders efficient search, comparison, and understanding of legal provisions. At the same time, for researchers and practitioners working in artificial intelligence and natural language processing, PDF-based corpora present substantial obstacles to the creation of high-quality training datasets. Extracting reliable, well-structured data from documents that were never designed for semantic processing requires extensive preprocessing, heuristics, and manual correction.

Importantly, this challenge is not limited to legislative documents. The widespread use of PDF can be observed across numerous private companies, public institutions, and non-governmental organisations. In many organisational workflows, PDF has become the final and often the only preserved representation of internal knowledge, regardless of the complexity or semantic richness of the original content.

As a consequence, the semantic knowledge held by document authors—who understand the meaning, structure, and intent of the content at the time of creation—is effectively lost at the moment of publication. Once converted into PDF, documents no longer "know" their own structure, and this knowledge cannot be recovered without significant effort. This loss becomes particularly critical when organisations attempt to introduce AI-based solutions for document analysis, automation, or knowledge extraction. Such efforts are inherently constrained by the need to extract structured information from a format that was never intended to support semantic representation.

The remainder of this paper is organized as follows. Section 2 provides an in-depth discussion of the limitations of the PDF format with respect to semantic representation and logical document structure, and motivates the need for document models that explicitly preserve meaning and structure. Section 3 describes a service developed within the CAISE project that parses PDF documents into a structured JSON representation, enabling further automated processing and analysis of document content.

2. Problem Statement

Despite decades of progress in digital document technologies, the PDF format remains the dominant medium for document publication across public administration, industry, and academia. Its popularity stems primarily from its ability to preserve visual fidelity and

provide a platform-independent representation of documents. As a result, PDF has become widely perceived as a reliable and final form of publication, particularly for official and archival purposes. However, this dominance comes at a significant cost to semantic integrity and machine interpretability.

From a document engineering perspective, PDF is fundamentally a presentation-oriented format. As described in foundational work on document structures, a document should be understood as a combination of content, structure, and presentation, with clear separation between these layers [4]. PDF violates this separation by encoding documents primarily as a sequence of positioned visual objects optimised for rendering rather than for semantic interpretation. Consequently, the logical organisation of content—such as hierarchical structure, functional roles of elements, and explicit relationships between document components—is not explicitly represented.

This design choice has direct implications for information extraction. Numerous studies in document analysis and document understanding demonstrate that recovering structure from PDF files requires reconstructing semantic relationships from layout cues such as spatial alignment, font attributes, and geometric proximity [5,6]. The existence of entire research directions and datasets devoted to layout-based document understanding underscores the fact that PDF documents do not natively encode the information required for reliable automated processing. Instead, structure must be inferred heuristically or learned statistically, making the extraction process inherently complex and error-prone.

The persistence of PDF as a publication format can be understood through the metaphor of digital paper. As noted in theoretical work on documents and digital media, many digital documents continue to replicate the affordances of paper rather than exploiting the capabilities of computational systems [7, 8]. PDF reinforces this paradigm by privileging appearance over meaning, thereby encouraging authors and institutions to treat documents as visual artefacts rather than as structured data resources. This approach obscures the semantic knowledge possessed by document authors at creation time and prevents its direct reuse in downstream processes.

In legal and governmental contexts, these limitations have particularly severe consequences. Modern approaches to digital legislation emphasize the need for machine-readable legal texts that support interoperability, versioning, and fine-grained referencing [9]. Nevertheless, many legislative systems—including the Polish legislative framework—have historically relied on PDF as the primary format for publishing legal acts. While visually accessible, such publications significantly constrain citizens' ability to search, analyze, and understand legal information, and they hinder the development of advanced legal information systems.

The challenges posed by PDF are further amplified in the context of artificial intelligence and natural language processing. High-quality training data require consistent structure, explicit annotations, and reliable textual representations. PDF-based corpora rarely meet these requirements. As a result, researchers must invest substantial effort in preprocessing, error correction, and manual validation before documents can be used for machine learning purposes. Recent advances in multimodal document understanding models, such as LayoutLM [5], explicitly acknowledge this limitation by combining visual layout information with textual content to compensate for the lack of native semantic structure.

Recent advances in intelligent document processing increasingly combine OCR, layout analysis, and multimodal deep learning approaches. Contemporary frameworks such as PaddleOCR and related vision-language document understanding systems demonstrate the growing importance of integrating visual and textual representations in large-scale document analysis pipelines [10, 11]. These developments further confirm that effective semantic processing of PDF documents requires reconstruction of structure that is not explicitly preserved within the PDF specification itself.

In addition to misconceptions about structure, PDF is often perceived as a secure and passive format, particularly in comparison to editable document types. This perception is misleading. The PDF specification allows for the inclusion of executable content, such as embedded JavaScript, interactive forms, and multimedia elements [1]. Although these features are rarely used in standard publishing workflows, their existence contradicts the widespread assumption that PDF documents are inherently inert or immune to security-related concerns.

Taken together, these factors reveal a fundamental mismatch between the continued reliance on PDF for document publication and the growing demand for semantic accessibility, automation, and AI-driven analysis. The problem addressed in this work arises precisely from this mismatch: critical knowledge is routinely encapsulated in a format that neither preserves document semantics nor supports efficient machine-based processing. As long as PDF remains the primary publication format, advanced information access and intelligent document processing will remain constrained by the need to reconstruct meaning from a representation designed exclusively for visual consumption.

3. PDF-to-JSON Service for Structured Document Extraction

In order to enable effective processing of information contained in PDF documents, a dedicated PDF-to-JSON service was designed and implemented. The service was designed to bridge the gap between presentation-oriented PDF files and the requirements of automated, machine-based document processing.

As discussed in the previous sections, PDF documents do not natively preserve logical or semantic structure. Instead, they encode content as visually positioned elements optimised for rendering, which significantly limits the direct reuse of document content in automated processing pipelines.

The proposed PDF-to-JSON service extracts textual content together with associated formatting and spatial metadata and transforms the extracted information into a structured JSON representation. This intermediate representation enables subsequent processing stages, including logical structure reconstruction, semantic annotation, and downstream NLP-based analysis.

Figure 1 presents the processing pipeline adopted in this work for transforming PDF documents from a presentation-oriented format into semantically usable representations. The pipeline explicitly separates three conceptual layers: the PDF input layer, the structured intermediate representation, and the semantic processing layer.

The service processes arbitrary text-based PDF documents through a multi-stage extraction pipeline. A PDF document provided as input is parsed to extract low-level content streams, including character sequences, font attributes, and geometric positioning. Based on this information, the service distinguishes between different types of document elements, including plain text blocks, tables, and images.

Textual content is extracted together with detailed formatting metadata, such as font size, font family, text orientation, and bounding boxes. Each extracted element is precisely localised within the page coordinate system, allowing its spatial relationships to other elements to be preserved. Non-textual elements, including embedded images and tabular structures, are identified and explicitly marked in the output representation.

All extracted data are transformed into a unified JSON structure shown in Figure 2, in which each document element is represented as an independent object enriched with content, metadata, and positional information. This design enables consistent handling of heterogeneous document components and supports further automated analysis without direct dependence on the original PDF format.

The output of the service is a structured JSON document that captures the content and layout characteristics of the source PDF. Each extracted element includes identifiers for page and block position, bounding box coordinates, content type, and detailed typographic attributes.

By preserving spatial localisation and formatting metadata, the JSON representation retains information that is essential for reconstructing higher-level document structures, such as paragraphs, headers, tables, or logical sections. At the same time, the representation remains independent of any specific rendering technology, making it suitable for integration with downstream processing components.

The JSON format serves as an intermediate representation between presentation-oriented PDF documents and semantically enriched document models. It provides a foundation for tasks such as information extraction, indexing, document comparison, and conversion to formats that explicitly encode logical and semantic structure.

The service is implemented as a RESTful API using the FastAPI framework, enabling straightforward integration with external systems and processing pipelines. The implementation relies on a set of well-established Python libraries for PDF parsing, data modelling, and numerical processing.

The service supports the extraction of textual content from PDF documents together with the identification and marking of structural and non-textual elements. Its functionality includes:

- extraction of character sequences and text blocks,
- identification and annotation of tabular content,
- detection and marking of embedded images,
- preservation of formatting attributes and spatial localisation.

This functional scope is designed to support subsequent stages of document understanding, including logical structure reconstruction and semantic annotation.

The quality of the service was evaluated using the F1 measure, which captures the balance between precision and recall for the supported extraction tasks. The evaluation methodology and the definition of individual metric components are described in a dedicated functional benchmark.

The benchmark corpus consisted predominantly of Polish legal documents, which typically exhibit a high degree of textual density and a lack of interactive elements, reflecting the characteristic features of legal texts published in Poland. This choice of evaluation material ensured that the assessment was aligned with real-world document processing scenarios in the legal domain.

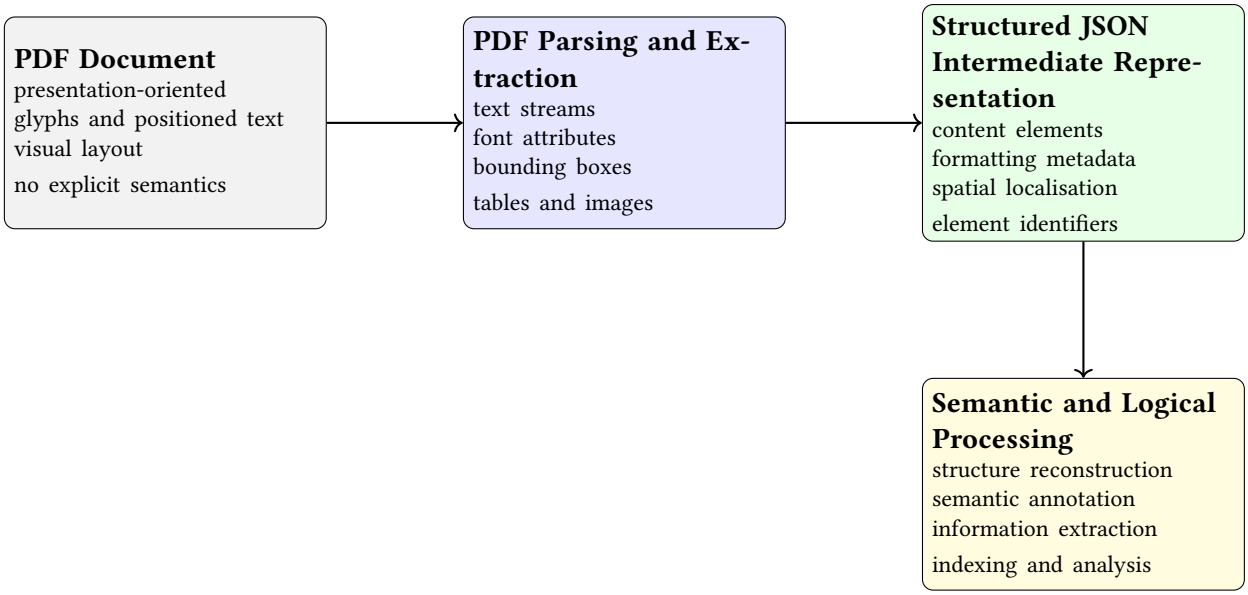

Figure 1: Processing pipeline for transforming presentation-oriented PDF documents into semantically usable representations.

Based on the conducted benchmark, the service achieved an F1 score of 0.858, exceeding the assumed acceptance threshold of F1 = 0.7 [12]. This result indicates that the service provides reliable extraction of document content and layout information suitable for further automated processing.

4. Summary

The achieved F1 score of 0.858 demonstrates that the proposed PDF-to-JSON service is sufficiently accurate to support effective information extraction from legal documents. In practical terms, this level of performance enables reliable downstream processing, including indexing, semantic annotation, and NLP-based analysis of legal texts published in PDF format.

At the same time, the obtained results highlight a fundamental limitation that does not originate from the extraction approach itself, but from the characteristics of the PDF format. The performance ceiling observed in the evaluation reflects the loss of semantic and logical information that occurs at the moment of document publication. Once a document is converted into PDF, its original structure, hierarchy, and functional roles are no longer explicitly available and must be reconstructed indirectly from visual layout cues.

From this perspective, an F1 score of 1.0 should be regarded not as a realistic target for PDF-based processing, but as a theoretical upper bound that could only be achieved if document semantics were preserved at the source. While numerous PDF parsing tools exist, they primarily focus on low-level text extraction or layout recovery and therefore operate within the same fundamental constraints imposed by the presentation-oriented nature of the format [13]. The need to infer structure from visual representations inevitably introduces ambiguity and error, even when advanced extraction and analysis techniques are applied.

It should also be noted that the current implementation of the service does not yet support the extraction of semantic information from certain specialised content types, such as mathematical or chemical formulas [14]. These structures are particularly challenging in PDF documents, as they are typically represented as complex visual compositions rather than as explicit symbolic expressions. Addressing such cases constitutes an additional and non-trivial research problem that lies beyond standard text and layout extraction.

The results presented in this work therefore support a broader conclusion: while engineering solutions such as the proposed PDF-to-JSON service can significantly mitigate the limitations of PDF-based workflows, they cannot fully compensate for the semantic degradation inherent in the format itself. Achieving truly lossless document understanding requires publication models that explicitly separate content, structure, and presentation and that preserve semantic information throughout the document lifecycle.

The rapid development of multimodal document understanding systems and OCR-oriented AI frameworks additionally highlights the limitations of traditional presentation-oriented document formats. Modern approaches increasingly rely on combining textual extraction, visual layout modelling, and semantic inference in order to compensate for the lack of explicit structural representation in PDF documents [10, 11]

In this sense, the proposed service should be viewed

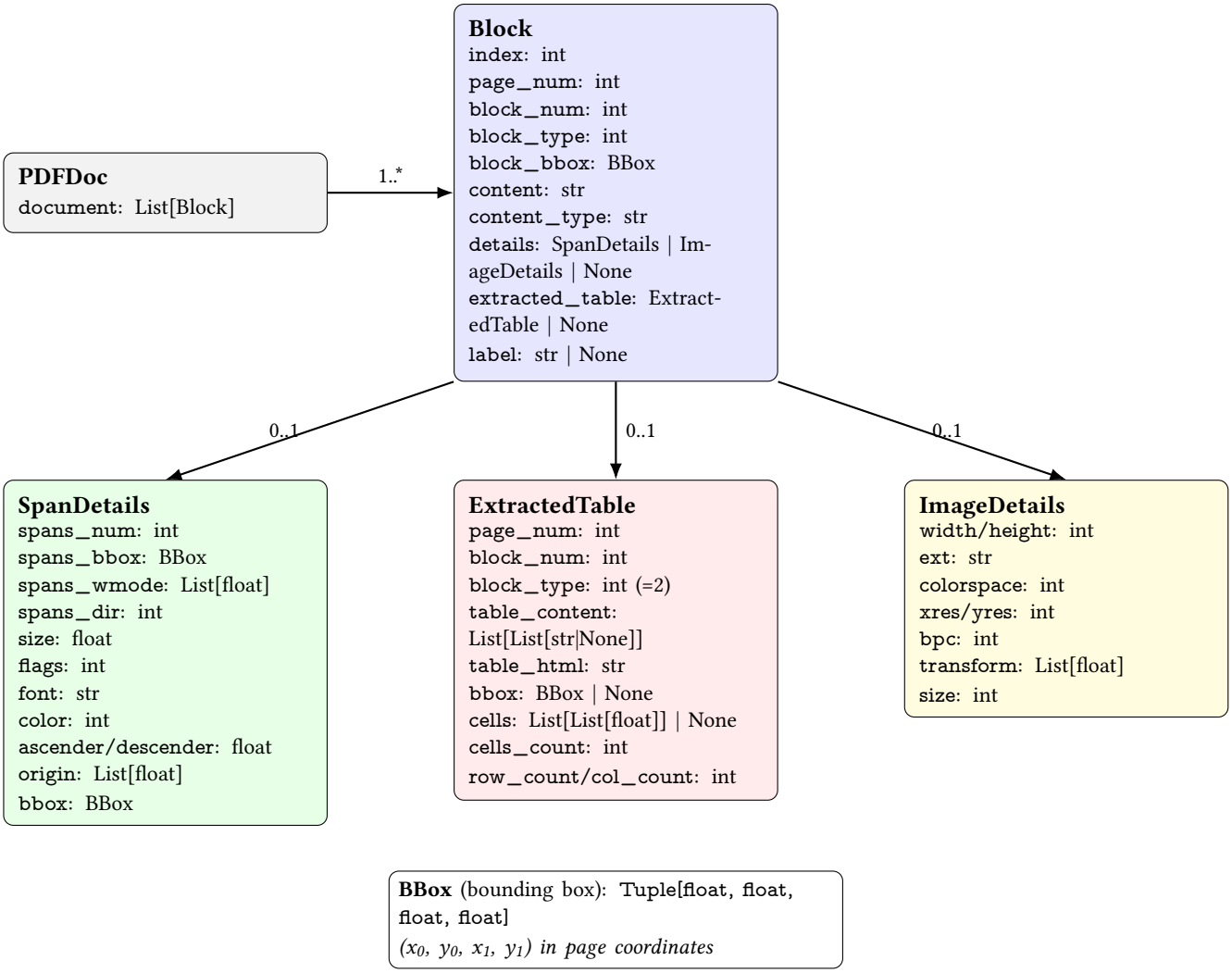

Figure 2: Schema of the structured JSON representation produced by the PDF-to-JSON service. The document is represented as a list of content blocks enriched with spatial metadata; block details are specialised for text spans or images, and optional table extraction is provided when applicable.

not only as a practical tool for processing legacy PDF documents, but also as empirical evidence of the structural constraints imposed by treating documents as digital paper. The findings reinforce the need for a transition towards semantically aware document formats and publication practices, particularly in domains such as law and public administration, where reliable access to structured information is essential.

Acknowledgements

The research was supported by project "Cloud Artificial Intelligence Service Engineering (CAISE) platform to create universal and smart services for various application areas", No. KPOD.05.10-IW.10-0005/24, as part of the European IPCEI-CIS program, financed by NRRP (National Recovery and Resilience Plan) funds.

References

[1] "PDF 1.7; Document management — Portable document format," standard, International Organization for Standardization, 07 2008.

[2] "Document management — electronic document file format for long-term preservation — part 1: Use of pdf 1.4 (pdf/a-1)," 2005.

[3] "Dziennik ustaw rzeczypospolitej polskiej." https://dziennikustaw.gov.pl. Official journal for the publication of legal acts in Poland.

[4] R. J. Glushko and T. Mcgrath, *Document Engineering: Analyzing and Designing Documents for Business Informatics and Web Services*. The MIT Press, August 2005.

[5] Y. Xu, Y. Xu, T. Lv, L. Cui, F. Wei, G. Wang, Y. Lu, D. Florencio, C. Zhang, W. Che, M. Zhang, and L. Zhou, "LayoutLMv2: Multimodal pre-training for visually-rich document understanding," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (C. Zong, F. Xia, W. Li, and R. Navigli, eds.), (Online), pp. 2579–2591, Association for Computational Linguistics, Aug. 2021.

[6] X. Zhong, J. Tang, and A. Jimeno-Yepes, "Publaynet: Largest dataset ever for document layout analysis," 09 2019.

[7] G. Nunberg, ed., *The Future of the Book*. Berkeley, CA: University of California Press, 1996.

[8] D. M. Levy, *Scrolling Forward: Making Sense of Documents in the Digital Age*. Arcade Publishing, 2001.

[9] F. Vitali, M. Palmirani, R. Sperberg, and V. Parisse, "Akoma Ntoso Version 1.0. Part 2: Specifications," standard, OASIS, August 2018.

[10] PaddlePaddle Authors, "PaddleOCR: Awesome Multilingual OCR Toolkits Based on PaddlePaddle." `https://github.com/PaddlePaddle/PaddleOCR`, 2026. Accessed: 2026-05.

[11] S. Li, A. Sadekar, N. Self, Y. Su, L. Andersland, M. Chaplin, A. Zhang, H. Yang, J. B. Henderson, K. R. Wigginton, L. C. Marr, T. M. Murali, N. Ramakrishnan, and V. Tech, "Exploring LLMs for scientific information extraction using the SciEx framework," *ArXiv*, vol. abs/2512.10004, 2025.

[12] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[13] "Apache tika: A toolkit for content analysis." `https://tika.apache.org`. Accessed: 2026-01.

[14] R. Zanibbi and D. Blostein, "Recognition and retrieval of mathematical expressions," *International Journal on Document Analysis and Recognition*, vol. 15, pp. 331–357, 2012.