Docora: A System for Interactive Knowledge Extraction and Visualization from Scientific PDFs

Dinh-Truong Do[1,2*], **Hoang-An Trieu**[1,2*], **Van-Thuy Phi**[1], **Le-Minh Nguyen**[2], **Yuji Matsumoto**[1]

[1]RIKEN Center for Advanced Intelligence Project, Tokyo, Japan
[2]Japan Advanced Institute of Science and Technology, Ishikawa, Japan
truong.do@riken.jp, an.trieu@riken.jp, thuy.phi@riken.jp, nguyenml@jaist.ac.jp, yuji.matsumoto@riken.jp

Abstract

Scientific research articles, typically distributed in PDF format, contain valuable knowledge but remain challenging to convert into structured datasets due to fragmented workflows that separate parsing, annotation, and visualization. Existing annotation platforms operate on plain text, which requires an additional PDF-to-text conversion step before annotation, while PDF parsing tools lack automated annotation suggestions. To bridge this gap, we introduce Docora, a system that unifies PDF parsing, automated annotation assistance, and multi-view visualization into a single interactive platform. Docora enables researchers to configure entity and relation schemas for any domain, automatically generates initial annotations using rule-based, model-based, or LLM-based extractors, and provides synchronized visualizations across PDF, text, and graph views. Users can refine annotations directly on the PDF canvas, ensuring consistency between document layout and structured representations. The system's source code is publicly available to facilitate further research and development.

Code — https://github.com/truongdo619/Docora

Introduction

Scientific literature remains the primary medium for disseminating new knowledge, with thousands of research articles—predominantly distributed in PDF format—published daily. These documents contain specialized terminology, complex conceptual relationships, and detailed descriptions of experimental or theoretical findings (Hanson et al. 2024; Stocker et al. 2025). Despite advances in information extraction, much of this knowledge remains effectively locked in formats optimized for human reading, posing significant challenges for processing and large-scale analysis (Nasar, Jaffry, and Malik 2018). For researchers aiming to construct structured datasets or knowledge graphs, the process often requires a labor-intensive workflow of PDF-to-text conversion, text normalization, and manual restoration of contextual information during annotation (Shen et al. 2022).

The lack of integrated systems that support the full extraction and annotation pipeline further complicates this task. Widely used annotation platforms such as Doccano (Nakayama et al. 2018), INCEpTION (Klie et al. 2018), Brat (Stenetorp et al. 2012), and Prodigy (Montani and Honnibal 2018) provide powerful interfaces for entity and relation annotation, but they operate only on preprocessed plain text. Conversely, PDF-direct parsing tools such as Grobid (Lopez 2009) and PolyMinder (Do et al. 2025) can extract textual content directly from PDFs but lack capabilities for automatic annotation suggestions or interactive visualization. As a result, researchers are forced to combine disparate tools, creating fragmented workflows that increase complexity, reduce efficiency, and introduce opportunities for error.

Tool	Domain	PDF	Suggest	Viz
Doccano	Any	✗	✗	Basic
INCEpTION	Any	✗	✓	Basic
Brat	Any	✗	✗	Basic
PolyMinder	Material	✓	✓	Basic
Grobid	Any	✓	✗	✗
Docora (ours)	Any	✓	✓	Rich

Table 1: Comparison of Docora with existing tools.

This gap highlights the need for a system that can process PDFs directly, extract entities and relations, and present them in context for interactive refinement (Nasar, Jaffry, and Malik 2018; Pawar, Bhattacharyya, and Palshikar 2023). To address this, we introduce Docora, a unified platform that combines parsing, annotation, and visualization into a seamless workflow. Unlike existing tools that either operate only on plain text or lack automated annotation suggestions (Table 1), Docora works on PDF layout while supporting domain-specific schemas, automated extractors, and synchronized multi-view visualizations. By consolidating these capabilities, Docora minimizes workflow fragmentation and enables efficient creation of context-preserving annotated corpora across diverse scientific domains.

System Design and Implementation

Docora is a domain-agnostic system for information extraction, annotation, and visualization that operates directly on PDF files. It unifies PDF content extraction, automated annotation, multi-view visualization, and interactive refinement into a single workflow (Figure 1). The frontend, built

*These authors contributed equally.

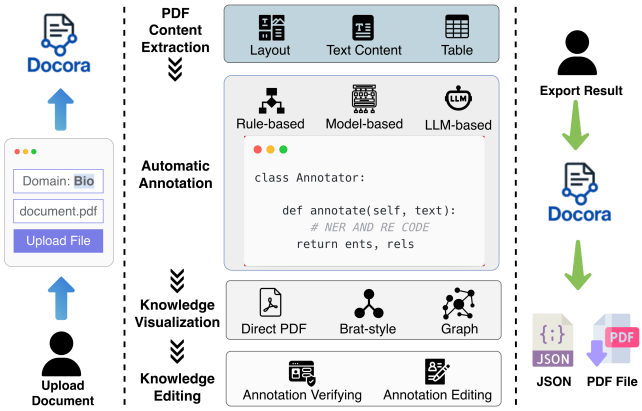

Figure 1: Docora architecture for PDF-to-knowledge.

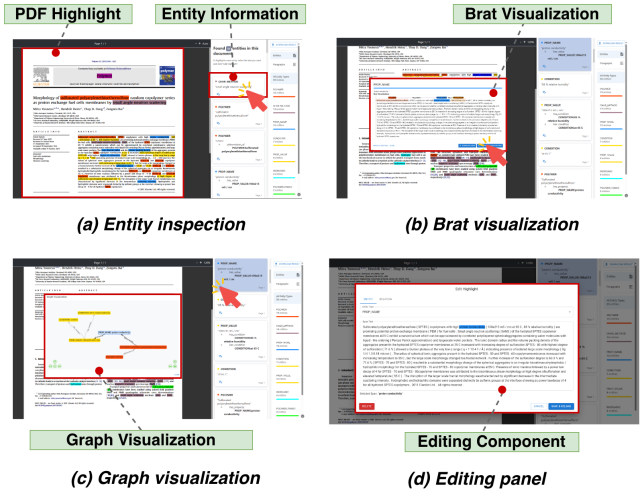

Figure 2: Screenshots of user interactions in Docora.

with React, HTML5, and CSS, provides an intuitive and responsive interface, while the backend coordinates PDF parsing and automated annotation, exposes APIs via FastAPI, and stores data in PostgreSQL.

PDF Content Extraction The pipeline begins with PDF content extraction, performing both layout and textual analysis. Scientific PDFs often include irrelevant information for knowledge extraction, such as publisher details, headers, or references. To filter these out, we first apply layout analysis using LayoutLMv3 (Huang et al. 2022), focusing only on the document's main content. Once the main layout is identified, we extract textual content and its positional information using PyMuPDF (Artifex Software 2025). This process yields raw text and corresponding coordinates, which are later used to highlight entities directly within the PDF file.

Automated Annotation Assistance Manual annotation of entities and relations is highly resource-intensive, especially when working with complex scientific or technical documents. To alleviate this burden, Docora incorporates a supporting annotation mechanism that automatically generates initial suggestions. These suggestions act as a first-pass analysis, providing users with a pre-labeled view of the document that can be verified and refined rather than created entirely from scratch. Depending on domain characteristics and the availability of annotated data, the system can employ rule-based matchers for domains with predictable patterns, fine-tuned Transformer models for domains with training corpora, large language models for low-resource or emerging domains, or a combination of them. At present, Docora supports three domains: material, biomedical, and legal. Importantly, its architecture is designed for extensibility, enabling administrators to add new domains with minimal effort by adhering to a template-based input–output convention, where raw text serves as input and the output is a structured set of entities and relations (Figure 1).

Integrated Multi-View Visualization Effective knowledge extraction requires preserving the contextual cues of the source document while also offering abstracted representations that aid interpretation. Docora achieves this through a suite of synchronized visualization modes that present information from complementary perspectives. Entities can be directly highlighted in the PDF view to maintain their original spatial and typographic context, allowing annotators to see exactly where information appears in the document's layout. At the same time, a side panel displays detailed metadata for each entity and its relations, enabling quick navigation and inspection without losing track of the surrounding content. For users who prefer a text-centered view, a Brat-style visualization (Stenetorp et al. 2012) renders the annotated text with labeled entities and relation arcs, making the logical structure explicit (Figure 2b). Finally, a graph-based visualization abstracts the document into a network of interconnected entities, revealing high-level patterns and relationships without the distraction of raw text (Figure 2c).

Interactive Editing and Refinement Since automated extraction is inevitably imperfect, Docora emphasizes rich and precise editing capabilities that allow users to validate, correct, and enhance the automatically generated annotations. Within the result interface, users can adjust entity boundaries, modify labels, create or delete relations, and correct misparsed text segments (Figure 2d). These edits are immediately reflected across all visualization modes and in the underlying structured data, guaranteeing that the document view, annotation layers, and export formats remain consistent at all times. Once annotation is complete, the enriched document can be exported either as a structured JSON file for computational processing or as an annotated PDF with embedded highlights for human-readable review.

Impact and Future Work

Docora unifies PDF parsing, automated annotation, and multi-view visualization into a single platform, reducing annotation effort, and enabling the creation of high-quality datasets. Moving forward, we aim to extend knowledge extraction to figures and tables and conduct user studies to evaluate effectiveness and usability. Feedback from researchers and practitioners will guide future development.

References

Artifex Software, I. 2025. PyMuPDF: Python bindings for MuPDF. https://pymupdf.readthedocs.io/. Version 1.24.9.

Do, T. D.; Trieu, A. H.; Phi, V.-T.; Nguyen, M. L.; and Matsumoto, Y. 2025. PolyMinder: A Support System for Entity Annotation and Relation Extraction in Polymer Science Documents. In Rambow, O.; Wanner, L.; Apidianaki, M.; Al-Khalifa, H.; Eugenio, B. D.; Schockaert, S.; Mather, B.; and Dras, M., eds., *Proceedings of the 31st International Conference on Computational Linguistics: System Demonstrations*, 1–8. Abu Dhabi, UAE: Association for Computational Linguistics.

Hanson, M. A.; Barreiro, P. G.; Crosetto, P.; and Brockington, D. 2024. The strain on scientific publishing. *Quantitative Science Studies*, 5(4): 823–843.

Huang, Y.; Lv, T.; Cui, L.; Lu, Y.; and Wei, F. 2022. LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, 4083–4091. New York, NY, USA: Association for Computing Machinery. ISBN 9781450392037.

Klie, J.-C.; Bugert, M.; Boullosa, B.; Eckart de Castilho, R.; and Gurevych, I. 2018. The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In Zhao, D., ed., *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, 5–9. Santa Fe, New Mexico: Association for Computational Linguistics.

Lopez, P. 2009. GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications. In Agosti, M.; Borbinha, J.; Kapidakis, S.; Papatheodorou, C.; and Tsakonas, G., eds., *Research and Advanced Technology for Digital Libraries*, 473–474. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-642-04346-8.

Montani, I.; and Honnibal, M. 2018. Prodigy: A Modern Annotation Tool for AI, Machine Learning and NLP. https://prodi.gy. Accessed: 2025-08-15.

Nakayama, H.; Kubo, T.; Kamura, J.; Taniguchi, Y.; and Liang, X. 2018. doccano: Text Annotation Tool for Human. Software available from https://github.com/doccano/doccano.

Nasar, Z.; Jaffry, S. W.; and Malik, M. K. 2018. Information extraction from scientific articles: a survey. *Scientometrics*, 117(3): 1931–1990.

Pawar, S.; Bhattacharyya, P.; and Palshikar, G. K. 2023. Techniques for Jointly Extracting Entities and Relations: A Survey. In Gelbukh, A., ed., *Computational Linguistics and Intelligent Text Processing*, 602–618. Cham: Springer Nature Switzerland. ISBN 978-3-031-24340-0.

Shen, Z.; Lo, K.; Wang, L. L.; Kuehl, B.; Weld, D. S.; and Downey, D. 2022. VILA: Improving Structured Content Extraction from Scientific PDFs Using Visual Layout Groups. *Transactions of the Association for Computational Linguistics*, 10: 376–392.

Stenetorp, P.; Pyysalo, S.; Topić, G.; Ohta, T.; Ananiadou, S.; and Tsujii, J. 2012. brat: a Web-based Tool for NLP-Assisted Text Annotation. In Segond, F., ed., *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 102–107. Avignon, France: Association for Computational Linguistics.

Stocker, M.; Snyder, L.; Anfuso, M.; Ludwig, O.; Thießen, F.; Farfar, K. E.; Haris, M.; Oelen, A.; and Jaradeh, M. Y. 2025. Rethinking the production and publication of machine-readable expressions of research findings. *Scientific Data*, 12(1): 677.