

# PolyMinder: A Support System for Entity Annotation and Relation Extraction in Polymer Science Documents

Dinh-Truong Do<sup>1,2</sup> Hoang-An Trieu<sup>1,2</sup> Van-Thuy Phi<sup>2</sup> Minh Le Nguyen<sup>1</sup> Yuji Matsumoto<sup>2</sup>

<sup>1</sup>Japan Advanced Institute of Science and Technology, Ishikawa, Japan

<sup>2</sup>RIKEN Center for Advanced Intelligence Project, Tokyo, Japan

## Objective & Contributions

### Objective

The objective is to provide an efficient, **web-based system** for **entity annotation** and **relation extraction** in **polymer science documents**, fostering large-scale annotation and organization of polymer-related information.

Our main contributions in this paper are as follow:

- Introducing PolyMinder, an annotation support system that extracts and visualizes key polymer-related information from scientific texts, allowing annotators to review and refine the data;
- Developing polymer-specific entity and relation extraction models utilizing state-of-the-art techniques tailored to the polymer science domain;
- Publicly releasing PolyMinder's source code to support further research and development: <https://github.com/truongdo619/PolyMinder>

## Introduction

**Polymers** drive advances in materials, manufacturing, biomedicine, and sustainability, creating a vast research corpus. Efficiently extracting polymer data is crucial yet challenging, requiring expertise. Automated NER systems using neural networks show promise but often overlook direct PDF handling and user-friendly visualizations, hindering error correction.

### Phase 1: Model Training



NER model training

RE model training

Polymer-specific NER and RE models

PolyNERE Corpus

### Phase 2: Inferencing

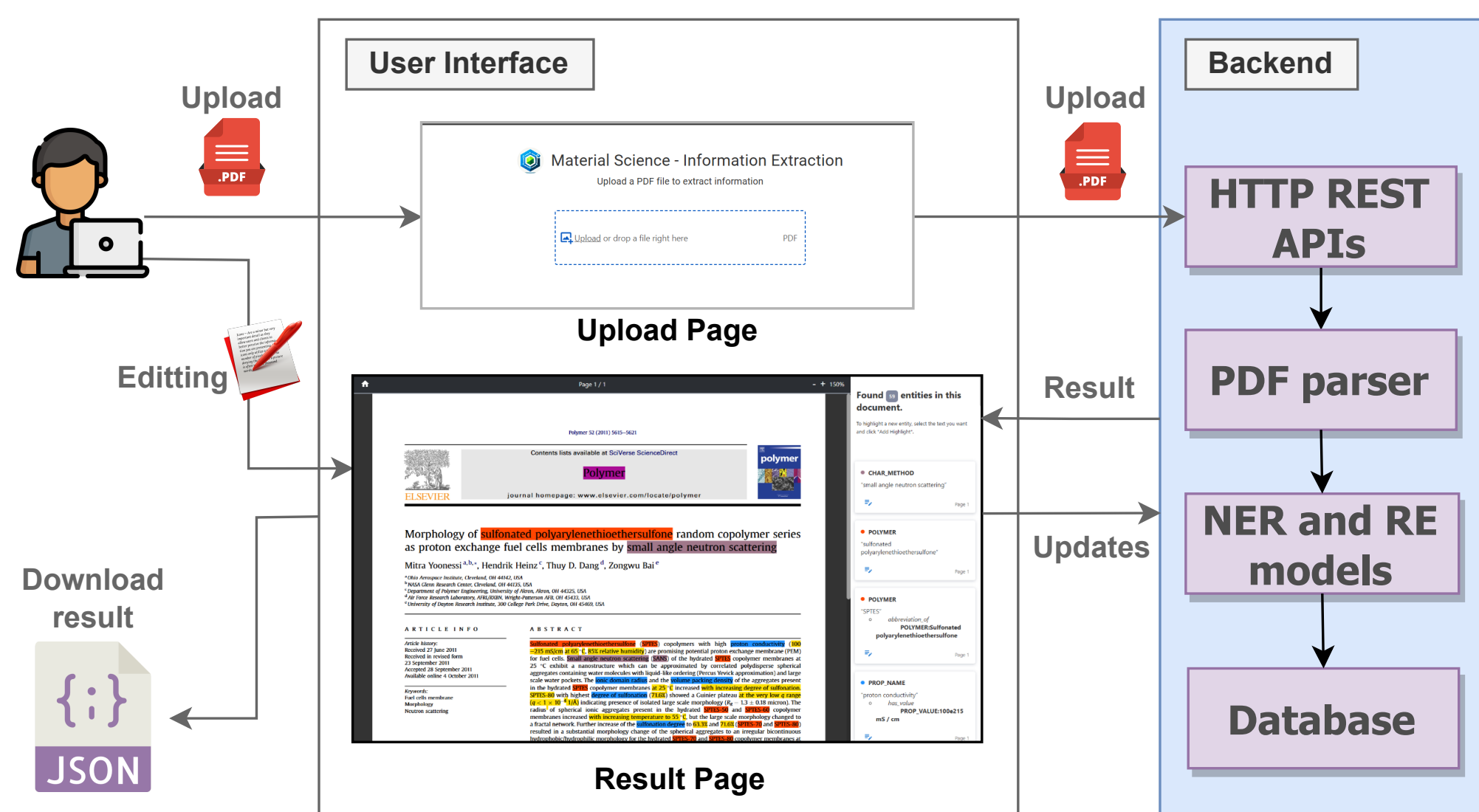


Figure 1: An overview of PolyMinder, showcasing its workflow from model training to entity/relation extraction and user interaction via the web-based interface.

**PolyMinder** is a polymer-domain annotation system addressing these limitations. Using advanced models from the PolyNERE corpus (Phi et al., 2024), it extracts entities and realtions from PDFs, provides intuitive visualizations, and supports user refinements, reducing manual annotation overhead.

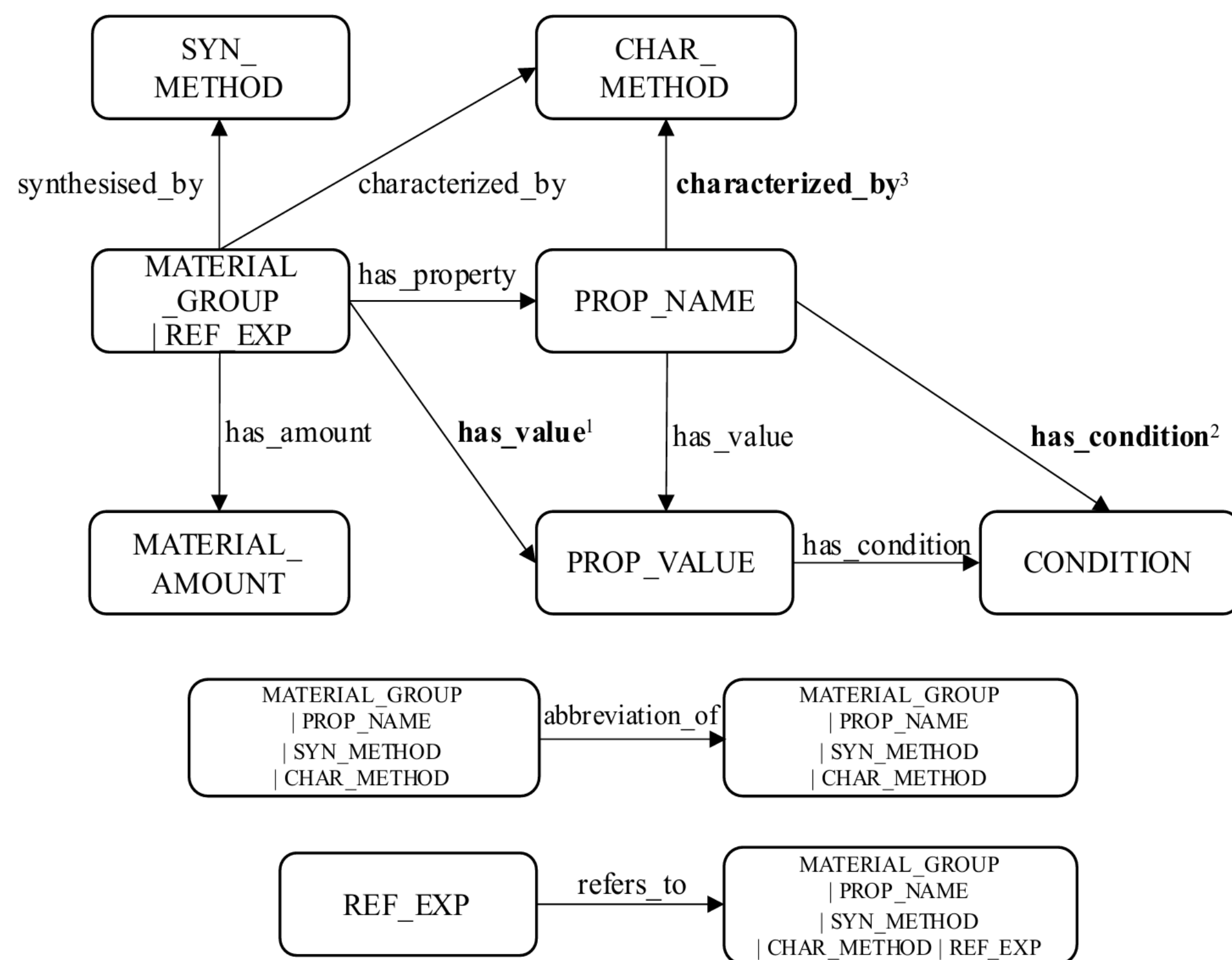
## Methodology

### Entity and Relation Extraction Backbone

PolyMinder extracts polymer-related entities with the W2NER model (Li et al., 2022), trained on the PolyNERE corpus (Phi et al., 2024) to handle flat, overlapped, and discontinuous mentions. The ATLOP model Zhou et al. (2021) then establishes paragraph-level relations between entities. Both models power PolyMinder's backend, producing JSON outputs. The tables below present the NER&RE model results, and Figure 2 visualizes the entities and their relations.

Method	Encoder	P	R	F1
W2NER	BERT-large	77.78	73.55	75.61
	SciBERT	74.89	75.67	75.28
Li et al. (2022)	MatSciBERT	78.05	76.53	77.28

Method	Encoder	P	R	F1
ATLOP	BERT-large	84.35	73.59	78.60
	SciBERT	83.59	81.60	82.58
Zhou et al. (2021)	MatSciBERT	83.99	82.49	83.23



- (1) When PROP\_NAME is missing or unclear
- (2) When PROP\_VALUE is missing or unclear
- (3) When MATERIAL\_GROUP is missing or unclear

Figure 2: Illustration of relations between different polymer-related entities within PolyMinder.

### PolyMinder System's Architecture

PolyMinder is a web app for extracting, visualizing, and annotating polymer data from PDFs. A Python backend (FastAPI) and React frontend deliver a seamless user experience. PyMuPDF parses PDFs, while NER and RE models identify key polymer entities and their relations, storing data via SQLAlchemy. RESTful APIs enable real-time, efficient data exchange.

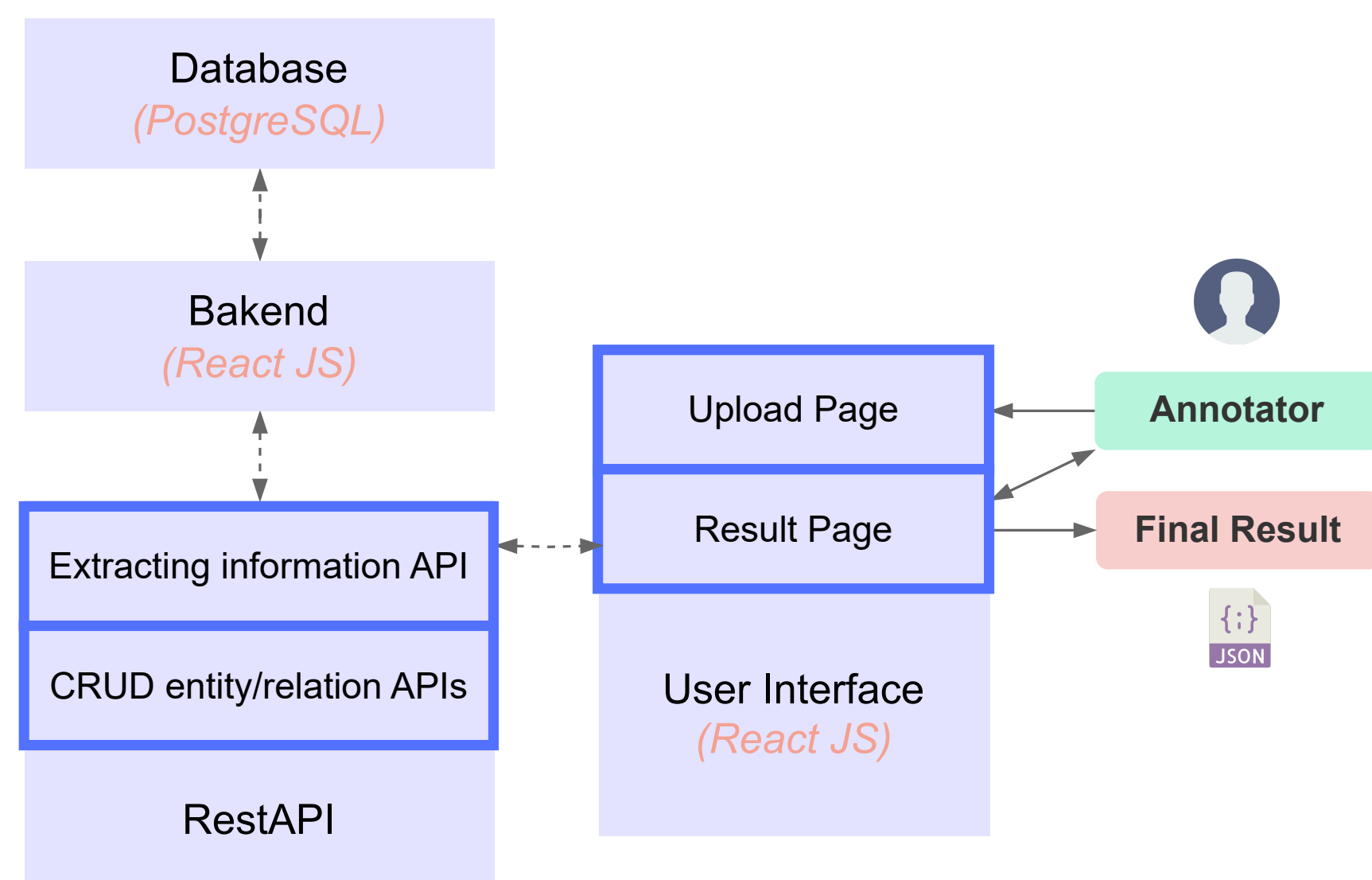


Figure 3: The architecture of PolyMinder system, detailing its backend, RestAPI, and frontend.

### Future Work

- Expand PolyNERE corpus for better generalization.
- Enhance PDF parsing for complex layouts.
- Adapt system to broader scientific domains.

## References

- Li, J., Fei, H., et al. (2022). Unified named entity recognition as word-word relation classification. In *Proceedings of the 36th conference on artificial intelligence*.
- Phi, V.-T., Teranishi, H., et al. (2024). PolyNERE: A novel ontology and corpus for named entity recognition and relation extraction in polymer science domain. In *Proceedings of Irec-coling 2024*.
- Zhou, W., Huang, K., et al. (2021). Document-level relation extraction with adaptive thresholding and localized context pooling. In *Proceedings of the aaai conference on artificial intelligence*.

### Web System



### Video

