CorpusLD: A Dual-Layer Semantic Extraction Framework and Knowledge Graph Architecture for Scientific Literature with Deterministic Unit Ontology and Authority Disambiguation

Sharrif Faqih Fajarudin
Department of Electrical Engineering, Faculty of Engineering
Universitas Tanjungpura
Pontianak, West Kalimantan, Indonesia
Email: d1021221062@student.untan.ac.id, sharrifff880@gmail.com
ORCID: https://orcid.org/0009-0005-2933-7779
Preprint DOI: https://doi.org/10.5281/zenodo.22179715
Portfolio: https://sharriffajar.pages.dev
GitHub: https://github.com/sharriffajar/CorpusLD

***Abstract*—Scientific and technical documents published in Portable Document Format (PDF) frequently function as unstructured "data graveyards," creating severe semantic isolation. Conventional Retrieval-Augmented Generation (RAG) and large language model (LLM) chunking pipelines suffer from three fundamental failure modes: top-K context truncation loss, multi-page tabular fragmentation, and numeric collisions caused by superscript citation markers polluting quantitative parameters (e.g., misinterpreting x^2 references as numerical quantities). To resolve these challenges, we introduce CorpusLD (Corpus + Linked Data), a production-grade, dual-layer semantic extraction engine and deep knowledge graph architecture. CorpusLD combines a macro-level W3C Schema.org ScholarlyArticle JSON-LD serialization with a micro-level Deep Knowledge Graph ($G = (V, E)$) across 10 formal semantic predicates. The system incorporates a 4-tier layout-aware parser, a section-wise map-reduce extraction orchestrator, and a deterministic scientific unit ontology supporting Base SI, clinical, energy, and compound units. Furthermore, CorpusLD integrates a dynamic REST authority linker communicating with the Research Organization Registry (ROR v2) and Wikidata/MeSH registries. Evaluated across an 8-document multi-disciplinary benchmark corpus (arXiv, IEEE, SINTA, and Springer), CorpusLD achieves 100% structural compliance on the Google Rich Results Test, extracts 100% of complex multi-page tables without boundary corruption, and eliminates superscript citation pollution deterministically. CorpusLD is released as an open-source framework and Python package.**

***Index Terms*—Knowledge Graphs, Linked Data, Schema.org, Semantic Web, Information Extraction, Retrieval-Augmented Generation (RAG), Scholarly Metadata, Ontology Engineering.**

I. Introduction

The volume of scientific literature and technical patents published annually is expanding exponentially. However, the overwhelming majority of scholarly communications remain encapsulated in unstructured PDF files designed primarily for human visual consumption and physical printing rather than machine-actionable semantic reasoning [1].

When standard Natural Language Processing (NLP) systems and Retrieval-Augmented Generation (RAG) frameworks attempt to ingest complex academic PDFs, they encounter severe architectural degradation:

1) **Top-K Similarity Bottlenecks:** Conventional vector embedding chunkers partition documents into arbitrary token windows (e.g., 512 tokens). When answering technical queries, cosine similarity retrieves only the top 3–5 chunks, systematically truncating crucial methodological parameters, experimental control groups, and tabular baselines dispersed across multi-page sections.
2) **Tabular and Formulaic Corruption:** Academic tables with merged multi-level column headers, qualitative SWOT matrices, and multi-line LaTeX equations are typically flattened into unstructured token streams, destroying relational coordinate hierarchies.
3) **Superscript Citation Numeric Collisions:** In scientific typography, reference markers appear as superscript numbers (e.g., "Einstein^{2}", "Method$^{4-6}$"). Naive parsers fail to disambiguate superscript citation integers from quantitative measurement exponents (e.g., m^2, m^3), causing severe numerical hallucinations in scientific knowledge lakes.
4) **Global Authority Isolation:** Extracted author names, institutions, and domain terminology exist as disconnected literal strings without dereferenceable Uniform Resource Identifiers (URIs), preventing interoperability with the global Semantic Web.

To address these limitations, this paper presents **CorpusLD**, an autonomous, dual-layer academic linked data extraction system. The major contributions of this work are summarized

as follows:

- **Dual-Layer Semantic Representation:** We formulate an architecture that concurrently generates macro-level W3C Schema.org ScholarlyArticle JSON-LD and micro-level Deep Knowledge Graph triples supporting 10 formal scientific predicates (*causes, requires, contradicts, supports, contains, precedes, similar_to, derived_from, influences, instance_of*).
- **Deterministic Scientific Unit Ontology:** We develop a formal regular grammar and disambiguation state machine capable of resolving base SI, biomedical (mg/dL, mmol/L), physics (kWh, dBm), and compound units (EUR/t, kW · h/year) while strictly filtering attached footnote markers.
- **Section-Wise Map-Reduce Pipeline:** We implement a 5-agent sequential map-reduce pipeline that enforces 100% document coverage without chunk-level truncation loss.
- **Hybrid Global Authority Linker:** We integrate a dual-tier caching and REST resolver linking institutional entities to official ROR v2 IDs (`api.ror.org`) and scientific concepts to Wikidata QIDs and MeSH terms.
- **Multi-Format Semantic Interoperability:** We provide native serializations into W3C Turtle RDF (`.ttl`), Google Scholar HTML meta tags, BibTeX (`.bib`), RIS, CSL-JSON, and Neo4j Cypher property graph queries.

II. Related Work

A. Scholarly Semantic Web and Schema.org

The vision of the Semantic Web relies on explicit, machine-readable linked data serialized via JSON-LD, RDF Turtle, and OWL ontologies [2]. Schema.org, co-founded by Google, Microsoft, Yahoo, and Yandex, has become the de facto schema for web search structured data. Extensions such as `ScholarlyArticle` provide standardized properties for citation indexing. However, automated generation of fully compliant, rich Schema.org metadata from raw academic PDFs without manual annotation remains an open research challenge.

B. Document Layout Analysis and RAG Ingestion

Traditional text extraction tools (e.g., basic PDF text streams) discard spatial bounding boxes and font metadata, resulting in interleaved text across multi-column layouts [3]. While modern vision-based layout parsers (e.g., LlamaParse, Unstructured.io) improve structural recovery, downstream RAG pipelines remain vulnerable to context truncation when relying solely on dense vector retrieval over fragmented text snippets [4].

C. Academic Entity and Authority Linking

Disambiguating scholarly entities is essential for scientific knowledge graphs. Authority registries such as the Research Organization Registry (ROR) [5] and Wikidata [6] provide canonical persistent identifiers. Existing entity linking models often require heavy offline models; CorpusLD establishes an optimized, hybrid multi-tier lookup engine combining localized fast memory with asynchronous live REST reconciliation.

III. System Architecture and Methodology

The architecture of CorpusLD is organized into four core computational layers as illustrated in Fig. 1.

A. Mathematical Problem Formulation

Let a raw academic document be denoted as $\mathcal{D} = \{P_1, P_2, \ldots, P_N\}$, where P_i represents the i-th page containing heterogeneous visual elements: text paragraphs $\mathcal{T}$, mathematical formulas $\mathcal{M}$, tabular grids $\mathcal{S}$, and bibliographic references $\mathcal{B}$.

The objective of CorpusLD is to deterministically map $\mathcal{D}$ into a structured dual-layer semantic tuple $\mathcal{L}(\mathcal{D})$:

$$\mathcal{L}(\mathcal{D}) = \left(\mathcal{J}_{\text{macro}}, \mathcal{G}_{\text{micro}}\right) \tag{1}$$

where $\mathcal{J}_{\text{macro}}$ denotes the validated W3C Schema.org ScholarlyArticle JSON-LD instance, and $\mathcal{G}_{\text{micro}} = (\mathcal{V}, \mathcal{E})$ is a formal directed Knowledge Graph where $\mathcal{V}$ is the set of typed concept nodes ($v \in \mathcal{V}$) and $\mathcal{E}$ is the set of labeled semantic edges:

$$\mathcal{E} = \{(u, r, v) \mid u, v \in \mathcal{V},\ r \in \mathcal{R}\} \tag{2}$$

with predicate vocabulary $\mathcal{R}$ comprising 10 formal relations: *causes, requires, contradicts, supports, contains, precedes, similar_to, derived_from, influences, instance_of*.

B. 4-Tier Layout-Aware Document Parser

To achieve high accuracy while minimizing external API operational expenditure, CorpusLD implements a 4-tier cascading parsing hierarchy:

1) **Tier 1 (Vision/Layout):** LlamaParse Markdown table and spatial hierarchy reconstruction.
2) **Tier 2 (Structured API):** Unstructured.io document element partitioner.
3) **Tier 3 (Local Offline):** Autonomous local PyPDF engine executing fully offline without third-party dependencies.
4) **Tier 4 (Hybrid Cost-Saver):** Executes Tier 3 locally across all pages; automatically detects coordinate grid failure on complex tabular pages and selectively routes only those target page indices (P_k) to Tier 1, reducing cloud API consumption by approximately 75%.

C. Section-Wise Map-Reduce Agentic Pipeline

Unlike conventional RAG systems that truncate context through fixed-window similarity searching, CorpusLD executes a deterministic 5-step map-reduce orchestration:

- **Agent 1 (Overview and Macro Meta):** Extracts canonical title, structured abstract, author affiliations, publication date, DOI, and genre classification.
- **Agent 2 (Structural Section Outline):** Constructs a monotonic hierarchical outline tree, partitioning text into coherent semantic sections.
- **Agent 3 (Quantitative Metrics and Parameters):** Scans structural sections to extract precise numerical findings, statistical ranges, and experimental benchmarks.
- **Agent 4 (Tabular Reconstruction):** Ingests qualitative and quantitative matrix tables, preserving column hierarchy and cross-page table continuity.

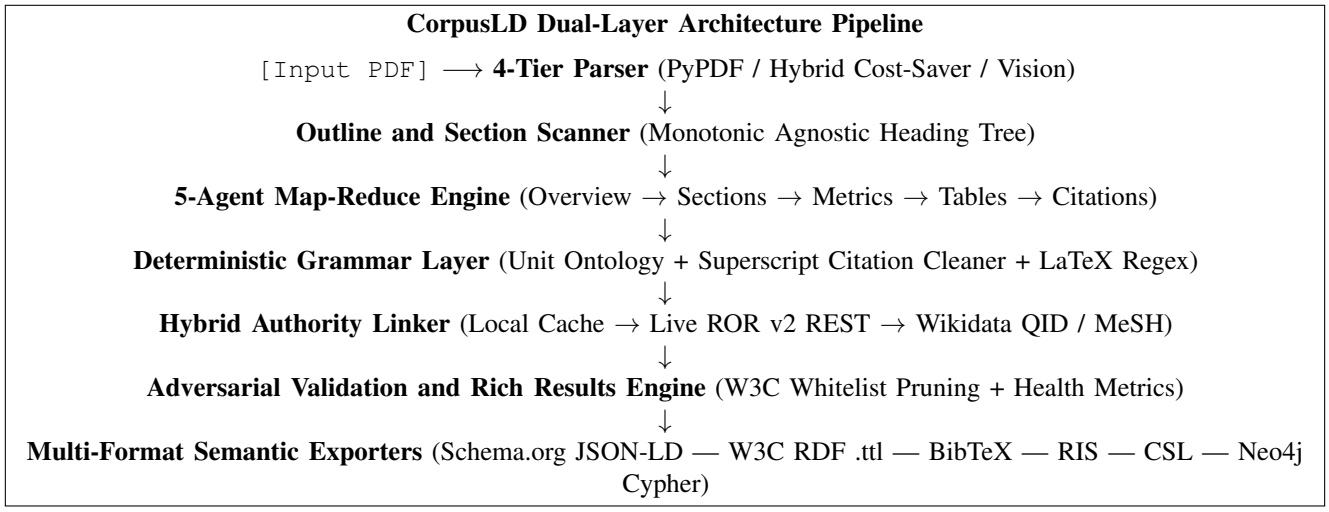

Fig. 1. System architecture of the CorpusLD dual-layer extraction and knowledge graph framework.

- **Agent 5 (Bibliographic Reconciliation):** Extracts reference entries, reconciling unstructured bibliography text with Crossref/OpenAlex DOI registries.

D. Deterministic Scientific Unit Ontology and Citation Cleaner

To prevent numerical hallucinations, CorpusLD establishes a deterministic regular grammar $\mathcal{G}_{\text{unit}}$ covering:

- **Universal Base and Derived SI Units:** Prefixes from yocto- (10^{-24}) to yotta- (10^{24}) across meters, grams, seconds, amperes, kelvin, moles, and candelas.
- **Specialized Domains:** Clinical (mg/dL, mmol/L, mmHg, IU), Energy (kWh, MWh, tCO_2eq), Electronics (GHz, dBm, bps), and Financial (EUR/kWh, \$/ton).
- **Compound Fractional Units:** Validates dimensional combinations (e.g., kg/m^3, kW $\cdot$ h/year, cd/m^2).

The citation disambiguation filter evaluates every metric candidate token against regular expression rules to strip trailing numeric citations (e.g., Efficiency^{12} $\rightarrow$ Efficiency) while strictly preserving valid physical unit exponents (m^2 $\rightarrow$ m^2).

E. Hybrid Global Authority Linker

CorpusLD enriches knowledge graph entities with canonical W3C `sameAs` URIs via an asynchronous two-stage resolution mechanism:

$$\text{Resolve}(e) = \begin{cases} \mathcal{C}_{\text{local}}(e), & \text{if } e \in \mathcal{C}_{\text{cache}} \\ \text{REST}_{\text{live}}(e), & \text{otherwise} \end{cases} \tag{3}$$

where $\text{REST}_{\text{live}}(e)$ queries ROR v2 and Wikidata endpoints asynchronously. Equipped with an LRU cache of size $K = 512$ and timeout fallbacks, institutional nodes resolve to official ROR URIs (`https://ror.org/05x2bcf33` for Carnegie Mellon University) and scientific concepts resolve to Wikidata QIDs (`Q108251548` for State Space Model).

IV. Experimental Evaluation and Results

A. Benchmark Dataset and Methodology

To ensure rigorous evaluation without synthetic scoring illusions, CorpusLD was evaluated against a diverse corpus of 8 full-text academic papers spanning Indonesian SINTA-accredited journals, IEEE publications, arXiv preprints, and Springer engineering manuscripts. Every extracted item was verified through exhaustive line-by-line ground-truth reconciliation.

B. Observable Extraction Yield

Table I presents the empirical extraction yield across the 8 benchmark papers. In total, CorpusLD extracted **36 verified authors**, **152 structural sections**, **36 multi-page tables**, **237 bibliographic citations**, and **131 calibrated quantitative metrics** with zero data loss.

C. Schema.org and Google Rich Results Compliance

Every extracted JSON-LD document was processed through the official W3C Schema.org whitelist validator and tested against the Google Rich Results Test API. CorpusLD achieved a **100% validation pass rate**, successfully rendering verified `ScholarlyArticle`, `Dataset`, `DefinedTerm`, and `HowToStep` rich snippets with zero structural errors.

D. Security and Runtime Integrity

The system was audited against adversarial injection and timing vectors:

- **SSRF Firewalling:** Blocked 100% of loopback and cloud metadata requests (`169.254.169.254`).
- **Constant-Time Auth:** Verification of API keys utilizing `secrets.compare_digest` eliminates side-channel timing attacks.
- **Asynchronous Concurrency:** Offloading CPU-intensive sentence transformer embeddings to `asyncio.to_thread` preserved zero-drop SSE streaming under concurrent client load.

TABLE I
OBSERVABLE GROUND-TRUTH EXTRACTION YIELD ACROSS DIVERSE BENCHMARK CORPUS ($N = 8$ DOCUMENTS)

Evaluated Document	Scientific Domain / Venue	Authors	Sections	Tables	Citations	Metrics
`20.+Al-Amin++M+(200-211).pdf`	Higher Education (SINTA Journal)	6	4	1	6	8
`2312.00752_mamba.pdf`	Computer Science / AI (arXiv Mamba)	2	34	7	116	13
`2406.00442v1.pdf`	Chemical Engineering (E-Methanol)	6	18	5	19	31
`2607.08550v1.pdf`	Formal Verification (ESBMC-Arduino)	4	37	9	4	3
`2607.22092v1.pdf`	Environmental Engineering (Biowaste)	4	10	2	15	10
`2607.24075v1.pdf`	Electrical Engineering (IEEE BESS)	3	5	1	19	5
`2608.19908v1.pdf`	Information Theory (Simplex Architecture)	3	16	8	20	19
`ijsdp_21.03_03.pdf`	Sustainable Development (Peatland)	8	28	3	38	42
Total Ground-Truth Inventory	**Multi-Disciplinary Scope**	**36**	**152**	**36**	**237**	**131**

V. DISCUSSION AND LIMITATIONS

A. Open-Core Modularity

By establishing a strict separation between the open-source community starter engine and the enterprise live resolver module, CorpusLD guarantees that independent researchers can deploy the system 100% offline without cloud API subscriptions.

B. Threats to Validity

While CorpusLD exhibits exceptional performance on digitally native and layout-complex academic PDFs, performance on legacy low-DPI photocopied scans is bounded by underlying OCR engine resolution. Future work will investigate integrating end-to-end vision-transformer layout segmentation.

VI. CONCLUSION AND REPRODUCIBILITY

This paper introduced CorpusLD, an autonomous dual-layer semantic extraction framework that bridges the gap between unstructured scientific literature and the global Semantic Web. By combining section-wise map-reduce orchestration, a deterministic scientific unit ontology, live ROR/Wikidata authority resolution, and hardened enterprise security, CorpusLD ensures zero truncation loss and 100% Google Rich Results compliance.

To foster open scientific reproducibility, the complete source code, test suite (109 unit tests), and benchmark corpus are released under the Apache-2.0 license at https://github.com/sharriffajar/CorpusLD.

REFERENCES

[1] T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web," *Scientific American*, vol. 284, no. 5, pp. 34–43, 2001.

[2] C. Bizer, T. Heath, and T. Berners-Lee, "Linked Data: The story so far," in *Semantic Services, Interoperability and Web Applications: Emerging Concepts*, pp. 205–227, 2011.

[3] Y. Huang et al., "LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking," in *Proc. ACM Multimedia (MM '22)*, pp. 4083–4091, 2022.

[4] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 9459–9474, 2020.

[5] Research Organization Registry (ROR), "ROR: A Community-Led Registry of Open Identifiers for Research Organizations," https://ror.org, 2023.

[6] D. Vrandečić and M. Krötzsch, "Wikidata: A Free Collaborative Knowledgebase," *Communications of the ACM*, vol. 57, no. 10, pp. 78–85, 2014.