Lawgistic AI

1. Project Overview

Lawgistic AI is an end-to-end legal data processing and intelligence platform designed to convert unstructured legal documents into clean, structured, machine-readable JSON for downstream AI applications such as search, RAG systems, analytics, and legal assistants.

The system focuses on **accuracy, determinism, and scale**, avoiding hallucination-prone methods and prioritizing traceable document parsing.

---

2. Problem Statement

Legal data exists in fragmented, inconsistent, and legacy formats (PDF, DOCX, DOC, RTF, scanned files). These documents:

- Lack structural consistency
- Contain noisy text, tables, and formatting artifacts
- Are difficult to query programmatically

Manual conversion is slow, error-prone, and not scalable.

Lawgistic AI solves this by automating high-volume, structure-preserving document conversion with failure tracking.

---

3. Core Objectives

- Convert heterogeneous legal documents into **standardized JSON**
- Preserve **section hierarchy and legal structure**
- Enable **AI-ready datasets** for legal search and reasoning
- Provide **conversion transparency** via failure reports

---

4. Supported Input Formats

- PDF (text-based)
- DOCX
- DOC

- RTF
- Scanned PDFs (handled separately where applicable)

---

5. System Architecture

5.1 File & OS Handling

- Batch-level file ingestion
- OS-safe file traversal
- Extension-based routing

5.2 Text Extraction Layer

Different extractors are used based on file type:

- **PDF**: Text extraction with layout tolerance
- **DOCX**: XML-based text parsing
- **DOC**: Legacy document parsing
- **RTF**: Rich text normalization

5.3 Table Extraction

- Tables parsed separately to avoid text corruption
- Structured rows retained where possible

5.4 Text Cleaning & Normalization

- Whitespace normalization
- Header/footer removal
- Noise filtering
- Regex-based cleanup rules

5.5 Section Extraction Engine

- Regex-driven section detection
- Hierarchy-aware parsing (Sections / Subsections)
- Deterministic rules (no AI guessing)

5.6 JSON Creation

Each document is converted into structured JSON containing:

- Metadata

- Section numbers
- Section titles (short, keyword-based)
- Full section descriptions

5.7 Failure & Audit Reporting

- Excel-based failure reports
- Conversion status tracking
- Transparent error categorization

---

6. Output Format (JSON – Simplified)

```
{
  "document_id": "unique_id",
  "title": "Document Title",
  "sections": [
    {
      "section_no": "1",
      "title": "Short descriptive heading",
      "description": "Full section text"
    }
  ]
}
```

---

7. Dataset Processing Summary

7.1 Overall Conversion Statistics

Dataset	Total Files	Converted	Not Converted
Lawgistic AI Files 2	4,868	3,053	1,815
Lawgistic 1	12,521	9,904	2,617
Material for Lawgistic.ai	3,498	2,843	655
Total	**20,887**	**15,800**	**5,087**

---

8. Segregated Legal Data Categories

	Total Data	Converted	Not Converted
1. Bills and legislation, Explore Us, Hansard, Notice Papers, Questions for written answers	563	132	431

2. Bougainville Legislation	3	2	1
3. Constitutional Instrument	4	4	0
4. Court related Materials	3	2	1
5. Custom Legislation	11	11	0
6. Government Gazettes	1772	1374	398
7. Historical Legal instruments- CONSOLIDATED LEGISLATION 1986	426	406	20
8. Historical Legal instruments- LAWS OF TERITORY OF PNG 1949- 1951	54	54	0
9. Other Material	64	48	16
10. Papua New Guinea Primary Materials- Historical Legal Instruments	116	116	0
11. Papua New Guinea Sessional Legislation	1292	1205	87
12. Parliament Related Materials	2	1	1
13. PNG Consolidated Legislation	60	60	0
14. PRIMARY MATERIAL	615	607	8
15. Primary materials	451	450	1
16. Recent Update	2	2	0
17. RECENT UPDATES	4	4	0
Total =	5442	4478	964

---

9. Key Design Principles

- **Deterministic parsing** (no hallucinated sections)
- **Explainable failures** (nothing silently dropped)
- **Scalable batch processing**
- **AI-readiness by design**

---

10. Current Limitations

- **Scanned document dependency**: All *not converted* files are primarily **scanned-image based documents**.
- **OCR not enabled** in the current pipeline by design, to avoid hallucination, low-confidence text, and legal inaccuracies.

- As a result, image-only PDFs without embedded text are skipped or marked as failed.
- Highly inconsistent legacy documents reduce conversion rate.
- Some tables lose semantic meaning after extraction.

---

11. Future Enhancements

- OCR + layout-aware parsing
- Vector embedding & FAISS indexing
- RAG-ready ingestion pipeline
- Legal citation linking
- Accuracy scoring per document

---

12. Use Cases

- Legal search engines
- AI legal assistants
- Regulatory intelligence platforms
- Government legal digitization
- Case law analytics

---

13. Conclusion

Lawgistic AI provides a **robust, auditable, and scalable foundation** for transforming raw legal documents into structured legal intelligence. It bridges the gap between legacy legal data and modern AI systems with precision and transparency.