Scope of Work – Parsing and Structuring Monitorul Oficial Part IV Data (Non-S.A. Category)

1. Context

Monitorul Oficial al României, Part IV (MoF4), publishes mandatory legal notices for companies operating in Romania. These documents include events such as company formation, registered office changes, and statutory decisions.

Our organization has collected all MoF4 issues from **2001 to 2025** (~8 million files) from two distinct sources:

- **SOURCE_1** (with an existing Python parser)
- **SOURCE_2** (requiring parser adaptation)

This project focuses on extracting and structuring company data for all legal forms **other than S.A.**.

---

2. Objectives

- Parse and structure MoF4 data from HTML into JSON according to a defined schema.
- Implement categorization and filtering for **non-S.A.** documents.
- Test two parsing approaches in parallel (heuristic vs. LLM structured output) for feasibility, scalability, and accuracy.
- Deliver a **viable, benchmarked solution** capable of processing at least 80% of non-S.A. category documents successfully within the performance constraints.

---

3. Scope of Work

Step 1 – HTML Parsing

- **Use the existing script** for SOURCE_1 HTML.
- Adapt script to SOURCE_2 HTML, extracting:
    - Each entry as a separate JSON object
    - Entry number and year
    - Company name
    - Additional metadata if available (CUI, legal type)

Step 2 – Categorization

- Classify parsed JSON entries into company types (S.R.L., S.A., P.F.A., etc.).
- Separate each category for potential separate processing flows.
- Retain only categories **other than S.A.** for the main task.

Step 3 – Data Extraction

- Define **data model** for companies other than S.A. (based on the provided example format).
- Implement two parallel extraction approaches:
    1. **Heuristic approach** – Regex, spaCy, NLTK, potentially modularized per document subtype.
    2. **LLM structured output** – Using PydanticAI structured output, with hallucination detection.

Step 4 – Feasibility & Benchmarking

- Measure:
  - Accuracy (manual + pattern-based verification)
  - Processing throughput
  - Robustness to structural variations
- Decide on a final approach based on feasibility and scalability.

---

4. Deliverables

- Adapted parsing scripts for both sources
- Categorization module
- Data model definition
- Heuristic parsing module(s)
- LLM structured output parsing module
- Benchmark report with success rate, scalability metrics, and hallucination detection results
- Final recommended parsing pipeline

---

5. Acceptance Criteria

- Processing ≥80% of non-S.A. category documents accurately
- Throughput: 3–4M documents processed in ≤3 days
- Validation via manual checks and deterministic patterns
- Clear documentation of code and process

---

6. Technical Constraints

- **Language:** Python
- **Libraries:** PydanticAI, spaCy, NLTK, Regex
- **Project Management:** uv
- **Version Control:** Git + GitHub
- **Infrastructure:** Local or server-based execution; parallelization encouraged
- **Timeframe:** 1 week total

---

7. Risks & Considerations

- Variations in HTML structure between sources
- Incomplete or malformed data entries
- Potential need for document-subtype-specific parsers
- LLM throughput limits

---

8. Conclusion and Recap

The goal is to parse and structure 24 years of legal notices for companies other than S.A. from *Monitorul Oficial, Part IV*.

The plan involves adapting existing scripts to handle two different data sources, categorizing the documents by company type, and then focusing specifically on the non-S.A. category.

A key component of this one-week project is a comparative analysis of two distinct data extraction methods:

1. A **heuristic approach** using traditional tools like Regex, spaCy, and NLTK.
2. An **LLM-based approach** using structured output to generate JSON directly from the text.

The final deliverable is a complete, benchmarked processing pipeline. Success is measured by its ability to accurately process at least **80%** of non-S.A. documents, with a throughput high enough to process up to **4 million files within a 3-day timeframe**. The results of the comparative test will determine the final recommended approach.