

On-Premise RAG App with data preprocessing pipeline

1. Functional Requirements

1.1 Frontend (Web UI)

- Development of a user-friendly web interface (e.g., with **Streamlit or React**).
- **Upload functionality** for documents (PDF) with acceptable speed.
- Display of a **chat interface** for user interaction.
- Provision of a **search function** to quickly find relevant information from uploaded documents.
- Implementation of a **preview function** to highlight relevant sections in uploaded documents.
- REUSE CURRENT ELEMENTS FROM PREVIOUS PROJECT

1.2 Document Processing (File Processing)

- **Extraction of relevant pages** from uploaded PDFs.
- **OCR analysis** to extract text from images or scanned documents.
- Storage of extracted texts including **positioning information** (bounding boxes) for later UI highlighting.

1.3 Data Preprocessing

- Conversion of extracted data into a model-compliant format for subsequent analysis.
- **Entity Recognition (NER)** to identify key terms (e.g., "Concrete Type," "Supplier").
- Storage of structured information in an appropriate database.

1.4 Data Storage

- Storage of extracted texts, metadata, and structured entities.
- **Option 1: Graph Database (e.g., Neo4j)** for managing project, document, and entity relationships.
- **Option 2: Vector Store (e.g., ChromaDB)** for semantic search based on embeddings.

1.5 Question-Answer System

- Capability for **semantic search** within a project.
- **Retrieval module** that extracts relevant text sections from the database.
- Optional: Integration of an **LLM** to generate answers based on extracted document information (e.g., Llama or Hugging Face models).
- Linking of **answers with original sources** and marking relevant passages.

1.6 User Interaction

- Provision of a **question-answer interface** for user queries. → existing one
- Session-based storage of conversations (UUID approach for sessions).

- Show references in PDF → see examples I provided in Git repositories

1.7 Deployment & Infrastructure

- Local execution on an **on-premise GPU server**.
-

2. System Architecture

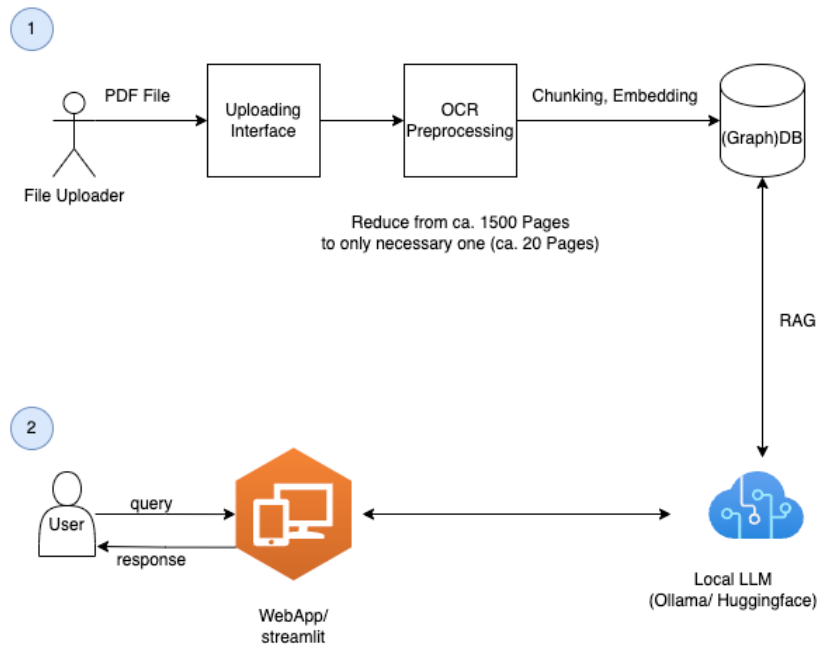
2.1 Process Workflow

1. **Upload** → PDF files are uploaded to the system.
2. **Preprocessing Step** → Identification of relevant pages, OCR processing, entity recognition, and storage of structured information.
3. **Database Storage** → Storage of extracted content in the **Vector Store or a GraphDB**.
4. **Frontend Integration** → The actual user interaction begins only after storage, as the UI accesses these structured data.
5. **Further development based on existing work** → Frontend work can build upon and extend the existing **Streamlit** demonstrator, leveraging already implemented functionalities.

2.2 Database Options

- **Graph Database** for relationships between projects, documents, and entities.
- **Vector Store** for semantic similarity search.
- **What are advantages/ disadvantages?**

Architecture



Comparison Feature with Original PDF (if possible)

User: Give me details about the storm Daniel

Assistant: Storm Daniel, also known as the Mediterranean cyclone, caused extreme flooding in Libya and other areas. It was a significant event in terms of loss of life, as mentioned in the provided text.

Type @ to mention a source...

Follow-up Send

Retrieved Sources

Search sources... Search

../data/1360_State-of-t... pdf
Relevance:

../data/1360_State-of-t... pdf
Relevance:

16 / 33

Extreme climate events

In 2023, many extreme climate events were reported across Africa. The continent was affected by heavy rainfall, floods, tropical cyclones, droughts, heatwaves, wildfires, and sandstorms. The extreme events in this section are described with respect to how they affected the different subregions.

FLOODS

EXTREME FLOODING IN LIBYA AND ELSEWHERE

In terms of loss of life, the most significant event was the Mediterranean cyclone, referred to locally as Storm Daniel, in September 2023. After affecting Greece, Bulgaria and Türkiye, the storm was slow-moving in the eastern Mediterranean for several days before the main rainbands impacted north-eastern Libya on 10 and 11 September. Extreme rainfall affected the coast and nearby mountains, with 414 mm falling in 24 hours at Al-Bayda on 10–11 September. The intense rainfall resulted in extreme flooding in the region. The most severe impacts were in the city of Derna (about 60 km east of Al-Bayda), where much of the central city was destroyed by flooding (Figure 8), exacerbated by the failure of two dams. At least 4 700 confirmed deaths in Libya have been attributed to the flooding with 8 000 still missing (as of 16 December 2023).¹

Tropical Cyclone Freddy, a long-lived cyclone in February and March 2023, formed off Australia's western coast and moved west across the Indian Ocean. It passed north of Mauritius and Réunion before making its first landfall on the east coast of Madagascar. Freddy re-intensified before making its second landfall in Mozambique. Although it dropped below cyclone intensity, it re-emerged over the Mozambique Channel and made its final landfall in Mozambique.

Figure 8. Flooded areas and destroyed buildings (red circles) in Derna, Libya on 10 September 2023
Source: European Union, Copernicus Sentinel image

References

I really like this Repo, (especially the view of the retrieved sources) so please check whether we can use components of it!

<https://github.com/Renumics/lexio/tree/main>