Flexible GovDoc Scanner with GFOSS

Basic Details:

- **Full Name:** Eftihis Drakakis
- **Email:** eftihisdrakakis@gmail.com
- **GitHub Username:** HixDr
- **First Language:** Greek
- **Location & Timezone:** Athens, Greece (EET / UTC+2)

Technical Skills:

- **JavaScript & Web Dev:** Worked with JS/TS on etherioparos.com.
- **Databases:** SQL experience from database systems courses.
- **AI:** Text preprocessing and feature extraction. Experience with Greek language technologies (GreekBERT)
- **Project-Specific:** Docker.

What is your motivation to take part in Google Summer of Code?

I see GSoC as an opportunity to work on a real-world open-source project while improving my technical skills and collaborating with experienced mentors. It's a great way to contribute to meaningful software, gain industry experience, and challenge myself with a complex project.

Why did you choose GFOSS?

GFOSS stands out to me because of its commitment to open data and public-interest technology. I'm passionate about open-source software and believe technology should be used to promote transparency and accessibility. GFOSS aligns perfectly with these values, making it the ideal organization for me to contribute to.

Why do you want to work on this particular project?

The Flexible GovDoc Scanner project is an exciting challenge at the intersection of web development, AI, and data processing. I'm particularly interested in working with Node.js, OCR, and scalable search solutions, and this project gives me a chance to apply these technologies to extract and structure valuable public data. It also has real-world applications in improving access to corporate information.

What are your expectations from us during and after successful completion of the program?

During the program, I hope to receive mentorship that helps me make the best technical choices and develop a high-quality solution. I want to improve not just my coding skills but also my ability to design and execute a structured project. After GSoC, I'd love to continue contributing to open-source projects, and if possible, stay involved in maintaining and improving this one.

Project Overview

Problem Statement:

In Greece, vital public company data is often locked in unstructured PDF files, making it challenging for citizens, researchers, and policymakers to access and analyze this information. The current state of these documents limits transparency and hinders efficient data use. The Flexible GovDoc Scanner project seeks to bridge this gap by transforming these PDFs into a structured, searchable database, thereby democratizing access to important corporate information.

What I am Making:

The Flexible GovDoc Scanner is an open-source tool designed to convert unstructured GEMI portal PDFs into a fully searchable database accessible via a REST API. The solution comprises the following steps:

- **Fetch**: Systematically retrieve documents from Greece’s official Open Data Portal, ensuring full compliance with legal and ethical standards.
- **Extract**: Use OCR (Tesseract) and advanced NLP tools (such as GreekBERT) to accurately extract metadata and text.
- **Organize:** Index and store the extracted data in OpenSearch, taking advantage of its built-in language analyzers for Greek text and fast querying capabilities.
- **Share:** Develop a robust REST API that allows users to query the database by various parameters such as company name, incorporation date, and key individuals, with features like pagination, filtering, and rate limiting.

How will it impact Open Technologies Alliance (GFOSS)?

The project enhances open data accessibility, directly supporting GFOSS’s goals of transparency and public-interest technology. By converting unstructured PDFs into a searchable database, it democratizes access to critical information and amplifies the societal value of open-source tools.

Why am I a good fit for this project?

My background in web development, AI, and database systems equips me with the skills needed to tackle this multifaceted project. I have practical experience with JavaScript and Docker, and my proficiency in Greek ensures accurate interpretation and processing of extracted text. I am self-motivated, detail-oriented, and committed to learning—qualities that will drive my success throughout this project. My experience and passion for open-source software make me an ideal candidate to contribute effectively to GFOSS.

What technologies will you be using?

- **Programming Languages:** JavaScript (Node.js)
- **Database:** Opensearch (why?)
- **AI & OCR:** Tesseract OCR, NLP tools
- **Infrastructure:** Docker, cloud services (e.g., AWS/DigitalOcean)

CLARIFICATION!!

We CANNOT crawl the main website (publicity.businessportal.gr):

- The main website’s robots.txt explicitly blocks all automated access (User-agent: * Disallow: /*).
- The Open Data Portal exists as the **lawful alternative** for accessing structured public data.

Timeline for Flexible GovDoc Scanner (GSoC 2025)

Community Bonding Period (May 8 – June 1)

Week 1 (May 8 - May 14):

- Collaborate with mentors to finalize project scope and Open Data Portal integration strategy.
- Review API documentation and dataset schemas; obtain the API key.

Week 2 (May 15 - May 21):

- Develop a strategy for effective text extraction from PDFs using OCR and NLP.
- Study Period: OCR/NLP optimizations and text extraction.

Week 3 (May 22 - May 28):

- Evaluate different database options with mentor guidance.
- Conduct a study on OpenSearch, focusing on indexing strategies and query optimizations.

Week 4 (May 29 - June 1):

- Finalize project architecture and tool selection.
- Set up the development environment (Docker, Node.js, search database).

- Study Period: Hands-on practice with building a simple REST API using Node.js and Express.
- Define milestones with mentors.

Coding Phase 1 (June 2 – July 14)
Objective: Implement Data Collection Mechanism
Week 5-6 (June 2 - June 15)

- Implement API requests to filter companies by specific dates and fetch PDF links.
- Filters are needed because the Open Data Portal provides a maximum result size of 200.
- Develop and test a PDF downloader capable of handling the result limitations of the Open Data Portal.
- Finalize a compact crawler that iterates through every company.

Week 7-8 (June 16 - June 29)

- Integrate OCR tools and advanced text extraction methods to retrieve detailed metadata from the PDFs.

Week 9 (June 30 - July 6)

- Begin integrating the extracted metadata into a search database using Opensearch.
- Conduct in-depth tests to ensure accurate data representation and efficient querying capabilities.

Week 10 (July 7 - July 14)

- Refine both the metadata extraction pipeline and database integration based on testing feedback.
- Complete the final testing of the entire pipeline (from API filtering to database indexing).

Coding Phase 2 (July 14 – August 25)
Objective: Develop and Deploy REST API for Metadata Search
Week 11 (July 14 - July 21):

- Configure the Node.js backend using a robust framework (e.g., Express).
- Establish project structure with version control and basic middleware for logging and error handling.
- Create initial endpoints (e.g., GET /metadata) to serve as a foundation for querying metadata.
- Implement basic routing and controller logic with placeholder responses.

Week 12 (July 22 - July 28):

- Integrate the chosen search database as the search backend.
- Define and implement data models/schemas to store extracted metadata such as GEMI ID, company name, representatives, and incorporation dates.
- Develop API endpoints to perform search queries on the database

Week 13 (July 29 - August 4):

- Implement pagination to handle large result sets.
- Improve API response times with optimizations.

Week 14 (August 5 - August 11):

- Add authentication and rate limiting to prevent abuse.
- Conduct API testing and validation with sample queries.

Week 15 (August 12 - August 18):

- Deploy the REST API on a cloud platform (e.g., AWS, DigitalOcean).
- Create API documentation and usage guides.

Week 16 (August 19 - August 25):

- Final testing and bug fixes.

Why Opensearch is the correct choice:

OpenSearch is the ideal choice for the Flexible GovDoc Scanner due to its focus on full-text search and rapid indexing of unstructured data. Key benefits include:

- **Language Support:** Built-in analyzers specifically tuned for Greek.
- **Scalability:** Efficient handling of large datasets.
- **Licensing:** Its Apache 2.0 license offers greater flexibility and lower costs compared to similar tools.
- **Integration:** Seamless integration with a Node.js environment through its modern REST API.

While alternatives like PostgreSQL or RavenDB focus on structured data storage, OpenSearch is purpose-built for the kind of unstructured, full-text search required for this project. I have evaluated potential challenges—such as indexing overhead—and will implement appropriate strategies to ensure optimal performance.