



ETH AI CENTER



United  
Nations

Practical Work - Project Report

# SemUN: A Semantics-Powered Search Platform for the United Nations' Digital Library

Clément Sicard  
csicard@ethz.ch  
D-INFK, ETH Zurich

*Supervised by:*

Dr. Menna El-Assady<sup>1</sup>, Dr. Sascha Langenbach<sup>1</sup>, Catherine Pysden, MSc.<sup>2</sup>

<sup>1</sup>ETH Zurich <sup>2</sup>United Nations

August 19, 2023

A faire

## Keywords:

*Natural Language Processing (NLP), Named Entity Recognition (NER), Graph databases, Frontend, Network Visualization*

## 1 Introduction

The United Nations Digital Library (UNDL) is a United Nations (UN) service that provides public access to a diverse range of UN documents: voting data, speeches, maps, and open access publications starting from 1979. All of these documents have been classified according to the UNBIS Thesaurus, a multilingual database of the controlled vocabulary used to describe UN documents and other materials in the Library's collection, with a more or less precise topic label.

The main idea of this project is to create an analysis platform, with a network visualization of documents from a subset of the documents in the UN Digital Library. The analytics platform includes a basic Named Entity Recognition (NER) system to extract mentioned entities from the documents. The visualization part will be a network visualization, leveraging the versatility of the structure of a graph to display the extracted insights. Both parts aim at improving the search of documents by implementing an analytics layer on top of the existing search engine from the digital library.

## 2 Motivation & Scope

In the short-term, the goal of this project is to provide a MVP of a potential future UN product, in close contact with UN staff to make it conform to their needs. Specifically,

this project will focus – as a first iteration – on Catherine Pysden’s suggestion, ”Women in Peacekeeping”.

The project’s long-term scope is to be used as a search engine for UN staff, member-state delegates, and members of the general public with an interest in UN topics. Potential future work could include extending the project to the whole UN Digital Library. It could, for instance, also suggest Thesaurus-compliant metadata for untagged documents to facilitate the work of Library staff.

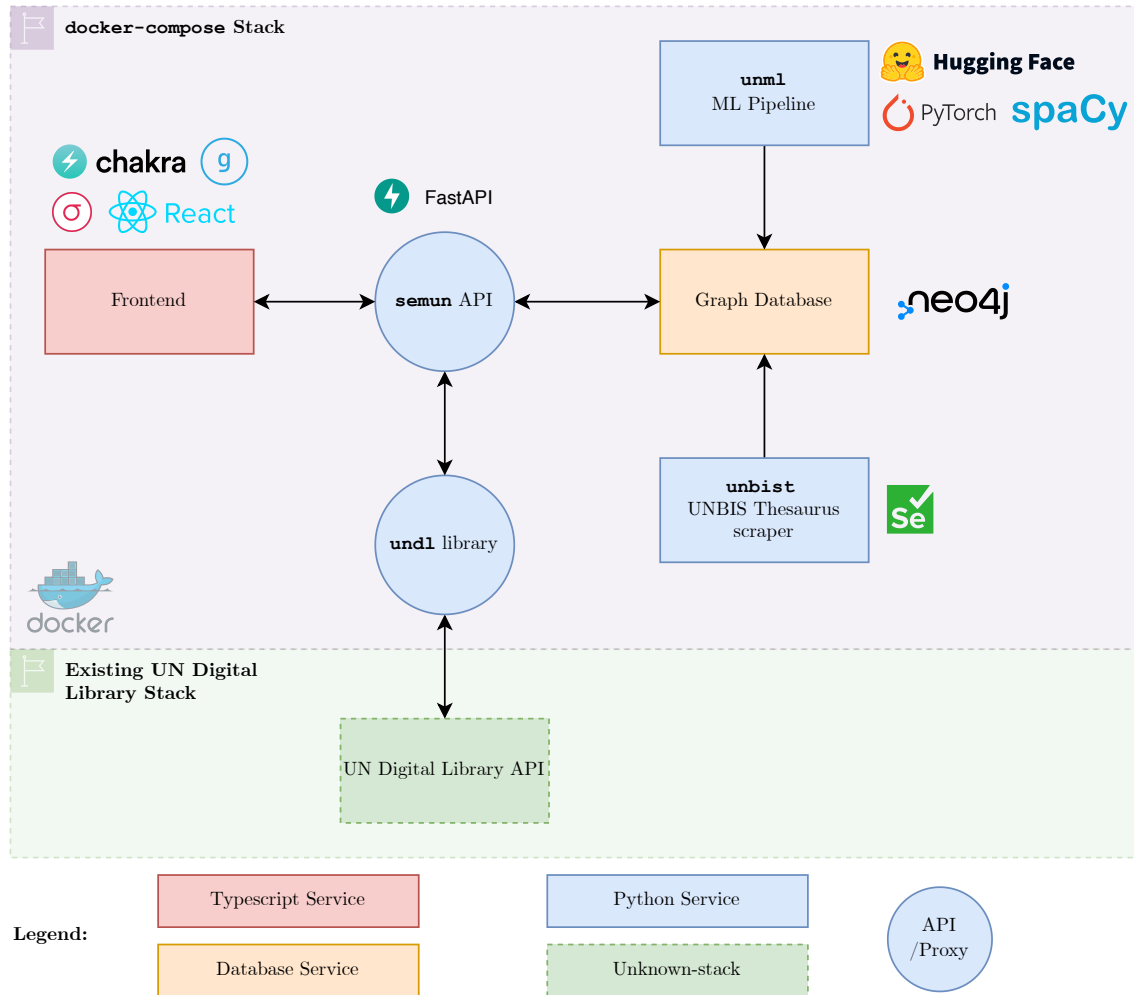











Figure 1: Final stack architecture

### 3 Final architecture

The final architecture is a full-stack architecture, from the database to the frontend, handed in as a Docker compose stack: **un-semun**.

The project consists of 7  Github repos:

-  **un-semun**: The main repository with the **docker-compose** stack declaration.
-  **un-semun-frontend**: The frontend.
-  **un-semun-api**: The API for the frontend.
-  **undl**: The code for **undl**, a Python wrapper around the UN Digital Library API.


-  **un-unbis-thesaurus-scraper**: A scraper for the UNBIS Thesaurus taxonomy website.
-  **un-ml-pipeline**: Machine learning pipeline for UNDL documents.
-  **un-semun-misc**: Diverse scripts used for the project.
-  **un-semun-paper**: The code for this paper.

This paper will go through each of them in detail except for the paper repository.

### 3.1 un-semun: The main repository

### 3.2 un-semun-frontend: A React & Sigma.js frontend

#### Tech stack of un-semun-frontend

I used React combined with Typescript for the UI framework, as well as Chakra UI for the UI components and **Sigma.js** via its React adapter **@react-sigma** for the network map. **graphology** was also used for graph manipulation in the frontend, mostly to iterate over graph elements to perform styling. The code is available here:  **un-semun-frontend**.

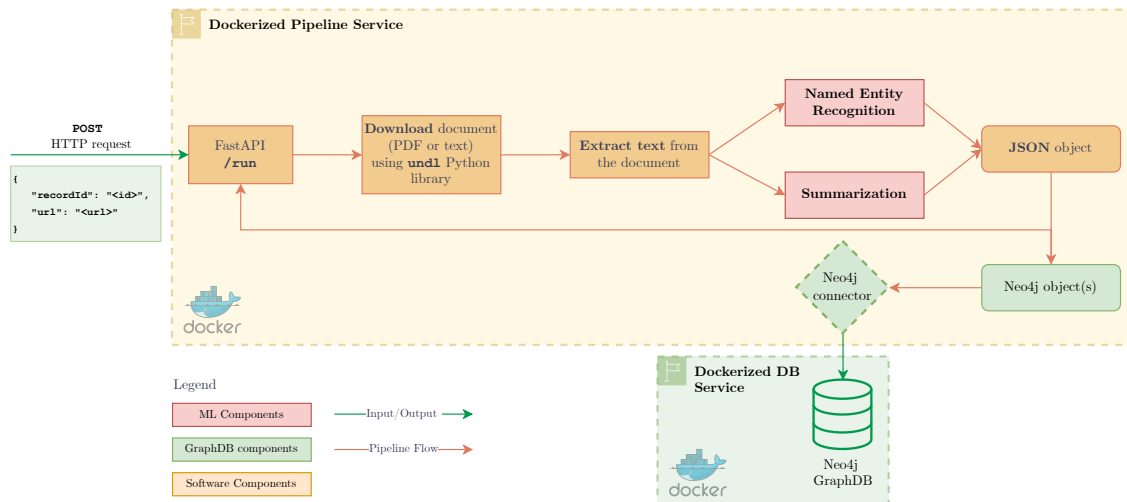


Figure 2: un-semun-frontend library: a dockerized machine learning pipeline for NER & summarization

### 3.3 un-semun-api: An API for un-semun-frontend using FastAPI

#### Tech stack of un-semun-api

### 3.4 undl: A Python library to wrap to the UN Digital Library API

#### Tech stack of undl

For the undl library, I used Python 3.10 with packages, with **requests** for the HTTP requests, **pandas** for the data manipulation, and **pydantic** for the data validation.

- Intermediary between frontend and Neo4j
- Proxying the UNDL API using undl library

### 3.5 un-unbis-thesaurus-scraper: the UNBIS Thesaurus scraper

Tech stack of un-unbis-thesaurus-scraper

### 3.6 unml: The machine learning pipeline

Tech stack of un-ml-pipeline

#### 3.6.1 Parts

Parts

#### 3.6.2 Models

Models

#### 3.6.3 API

API

Connection to Neo4j

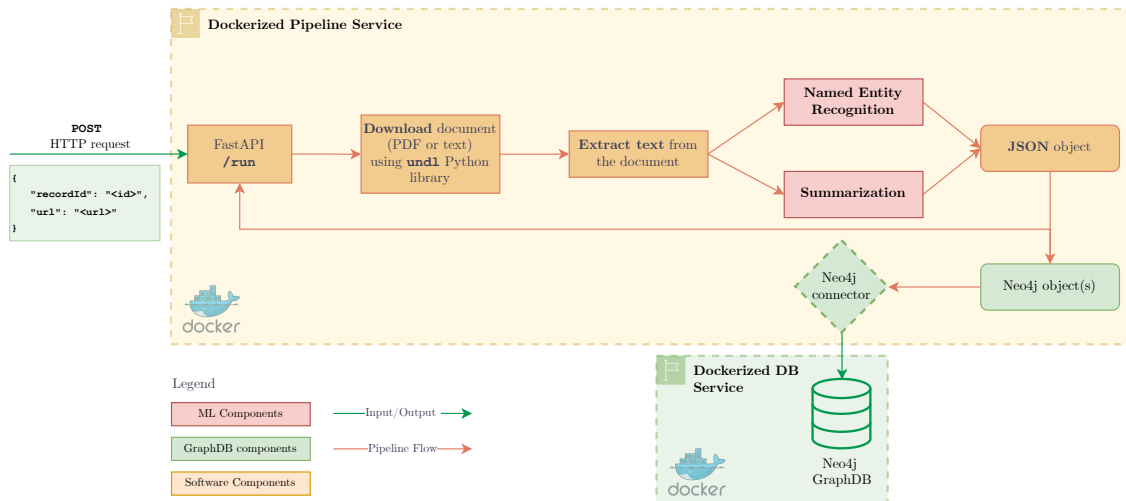


Figure 3: `unml` library: a dockerized machine learning pipeline for NER & summarization

### 3.7 Neo4j graph database

#### 3.7.1 Types of nodes

Types of nodes

#### 3.7.2 Types of relationships

Types of relationships

### 3.8 un-semun-misc

### 3.9 General tech stack notes

For all the Python components, the dependencies are managed using `poetry`. They are also all dockerized, and the whole stack is orchestrated using `docker-compose`.

## **4 Limitations**

The right of Figure 1 shows a simplified schema of the current UN Digital Library infrastructure.

## **5 Discussion & future work**

## **Conclusion**