Building the PDN2RDF MVP pipeline

Offline pipeline:

1- Build the dataset:
   a. Data extraction from ENA for the *priority pathogens* as defined by PDN. For the MVP, the data extracted focused on: ['run_accession', 'experiment_title', 'tax_id', 'country', 'description'].
   b. Obtain pathogen names from UniProt based on taxon IDs.
2- Define a schema profile to represent the extracted data using the InfectiousDisease class from https://schema.org/. The properties of the schema profile were:
   a. InfectiousDiseaseClass: Graph augmentation using a tool-binded LLM which queries UniProt using the pathogen name to obtain the type of pathogen (Virus, Fungus, Helminth, Bacteria, Prion, or Protozoa).
   b. run accession: obtained from ENA dataset.
   c. experiment title: obtained from ENA dataset.
   d. taxon id: obtained from ENA dataset.
   e. country: obtained from ENA dataset.
   f. associated disease: obtained from NCBI.
3- Populate the RDF graph based on the extracted data and the built schema profile.
4- Curate examples of Natural Language Questions with associated SPARQL Queries to enhance the performance of an LLM.

Online pipeline (based on ExpasyGPT):

1- LLM takes as input Natural Language Questions and generates a SPARQL query tested against the offline pipeline.
2- The offline pipeline returns a SPARQL query.
3- The LLM summarizes the results.