Assignment 1

Building a Knowledge Graph for Semantic Interoperability

Course: KEN 4256 – Knowledge Graphs
Due date: February 13, 2026 (23:59 CET)

Learning Objectives

By completing this assignment, you will learn to:

- Design a semantic data model for food-related datasets
- Construct a Knowledge Graph (KG) using RDF and existing vocabularies
- Integrate structured, unstructured, and external knowledge sources
- Query a KG using SPARQL to answer meaningful food and nutrition questions

Description

In this assignment, you will build a Knowledge Graph (KG) using RDFLib (Python) that integrates multiple food-related datasets. The goal is to integrate recipes, nutrition, ingredients, restaurants, cuisines and reviews, and to execute complex queries using SPARQL.

Data Sources

1. **Structured data**
    - **Recipes (CSV):** Recipes with details about ingredients, food types, preparation times, and instructions.
    - **Restaurants (CSV):** Information on restaurants, including location, cuisine type, price range, and rating.
    - **Nutrition (CSV):** Information on the nutrition of various recipes and food types.
2. **Unstructured data**
    - **Reviews (TXT):** Textual data discussing recipes and personal opinions.
3. **External KGs**
    - Wikidata

Tasks

Task 1: Schema Design (2 points)

Objective: Define a schema that integrates structured data and unstructured data.

1.1 Identify and define the main classes and properties across the provided datasets in a cohesive schema. Focus on properties that enable integration across datasets and support the SPARQL queries in Task 4. You are not required to include all columns from the original datasets in your schema.

Hint: The main classes include Recipe, Nutrition, Restaurant, and Review. Clearly indicate how these classes are connected.

1.2 Reuse Schema.org vocabularies and create a diagram showing classes and relationships between them.

1.3 Write a python code to represent the vocabulary in **RDF(S)** using Turtle syntax.

Task 2: KG Construction from Structured Data (1.5 points)

Objective: Programmatically convert structured data into RDF format.

2.1 Convert the Recipes, Restaurants, and Nutrition datasets into RDF triples. Assign unique URIs using the base namespace http://kg-course.io/food-nutrition/ and add rdf:type statements to all entities using the schema from Task 1.

Hint 1: To manage computational complexity and avoid scalability issues, you may restrict the RDF conversion to the first 10,000 rows of each dataset.

Hin 2: Link Recipe data and Restaurant data by string match on keywords and cuisines

---

Task 3: Enrich the KG with unstructured data and external KGs (3.5 Points)

Objective: Expand the constructed KG from Task 2 by integrating information from unstructured data and Wikidata.

3.1 Extract knowledge from unstructured data – use NLP tools (e.g., spaCy) to extract ingredient mentions and nutritional references and structure them in RDF.

(Hint: This task can be computationally resource-intensive. To ensure feasibility, limit the analysis to a maximum of 1,000 reviews)

3.2 Perform sentiment analysis on reviews (e.g., positive, neutral, negative), assign confidence scores, and link the results in the KG.

3.3 Link the extracted ingredients to Wikidata.

3.4 Integrate external KGs – query Wikidata to retrieve recipe-cuisine relationships and enrich the KG.

Hint: This task can be computationally resource-intensive. To ensure feasibility, limit the analysis to a maximum of 1,000 recipes

---

Task 4: SPARQL queries (8 Points)

Objective: write and execute SPARQL queries to answer questions about food and nutrition.

4.1 **(0.5 Point)** Find recipes that are risky for individuals allergic to mango.

- *Hint: dataset: recipes (based on RecipeIngredientParts)*

4.2 **(0.5 Point)** List all pies tagged as 'healthy' with a cooking time of less than 2 hours.

- *Hint: dataset: recipes (based on keyword, CookTime)*

4.3 **(0.5 Point)** List all restaurants in New Delhi that serve Chinese cuisine and offer online delivery.

- *Hint: dataset: restaurants*

4.4 **(0.5 Point)** Find the average cost of two dining at restaurants in Davenport that serve Asian food.

- *Hint: datasets: restaurants, Asian food includes Indian, sushi, Asian, Chinese, and Thai Cuisines.*

4.5 **(1 Point)** Recommend the top 5 desserts published after 2000 that are labeled as ‘Easy’ and low in calories(< 300), along with their images.

- *Hint: datasets: recipes+nutrition*

4.6 **(1 Point)** Identify the top 10 highly-rated beverages, including their preparation time and sugar content.

- *Hint: datasets: recipes+reviews+nutrition, RecipeCategory, requires NLP-based sentiment scoring*

4.7 **(2 Points)** Identify the highest-rated recipes containing protein-rich (> 20) ingredients and check if their corresponding cuisines are commonly available in the USA restaurants.

- *Hint: datasets: recipes+restaurants+nutrition+reviews*

4.8 **(2 Points)** Find the top 5 healthiest recipes based on a nutrition density score (NDS), along with their average sentiment from reviews. Check if restaurants serve cuisines linked to these recipes and retrieve their aggregate restaurant ratings.

To calculate an NDS, you will incorporate weighting factors based on dietary importance, i.e., high protein, high fiber, low sugar.

$$NDS = (w_p \times Protein(g)) + (w_f \times Fiber(g)) - (w_s \times Sugar(g))$$

Where:

- $w_p = 1.0$ (Protein is essential for muscle and metabolism)
- $w_f = 1.5$ (Fiber is crucial for digestion and satiety)
- $w_s = 2.0$ (Excess sugar contributes to metabolic issues, hence higher penalty)
- *Hint: datasets: recipes+nutrition+reviews+restaurants*

Deliverables

1. **Technical report (Max 5 Pages)**
    - Explain the methodology for each task, including schema development, vocabulary reuse, RDF conversion, entity linking, and sentiment analysis.
    - Include a diagram that represents the classes and properties used in the KG.
    - Summarize your SPARQL queries and key results, with explanations for design choices.
2. **Code files**
    - Python scripts (.py or .ipynb) for KG construction, integration, and querying, with clear documentation, in ready-to-run condition (all input file paths referenced correctly for example).
3. **Data outputs**
    - RDF(S) vocabulary and KG files in Turtle or N-Triples format.
        i. 'KEN4256-structured-KG-<team id>.ttl' containing KG constructed using structured data
        ii. 'KEN4256-unstructured-KG-<team id>.ttl' containing KG constructed using unstructured data

        iii. ‘KEN4256-integrated-KG-<team id>.ttl’ containing KG constructed using structured and unstructured data
    - SPARQL query results in `.txt` format.
4. **Contribution and reflection**
    - Describe team member contributions.
    - Reflect on the role of LLMs in your work, disclosing their use and impact.

    - Cite resources that informed the solution to this assignment.

Hints and Resources

- Refer to publicly available vocabularies like Schema.org, FoodOn, and Wikidata.
- Explore NLP tools like spaCy for entity extraction and Vader for sentiment analysis.

Reading Materials

- RDFLib documentation
- Easily deploy a SPARQL endpoint locally for your RDFlib graph: rdflib-endpoint ⭐
- TRANSFORMATION FROM SEMANTIC DATA MODEL TO RDF
- Data modelling with RDF: a tutorial
- Build a medium size KG from a CSV dataset
- Structured prompt interrogation and recursive extraction of semantics (SPIRES): A method for populating knowledge bases using zero-shot learning
- Introduction to OntoGPT
- Food Recipe Template
- Lecture 2 slides (Generate RDF graph from structured data)
- Lecture 3 slides (KG construction from unstructured data)
- Lab 2-5 materials

Questions and Comments

Prof. Dr. Michel Dumontier, michel.dumontier@maastrichtuniversity.nl

Maryam Mohammadi, m.mohammadi@maastrichtuniversity.nl