Knowledge Graph Creation

Task 1: Create a Knowledge Graph

The movie ontology is used as an initial ontology for this project. It can be found in the *input/graph/movie.ttl* file.

Since datasets used to populate the initial ontology contained more information, additional entities and properties (object and data) are added. Those are:

- **Location** – a new entity that will hold all the countries and their respective information
- **capital** – data property of Location, used to store the capital city of the country
- **currency** – data property used to store the currency of the country
- **code** – data property used to store the country code
- **locatedIn** – object property of Movie, used to store the country in which the movie is released
- **locatedIn** – object property of Film Studio, used to store where is the film studio located
- **releaseYear** – data property of Movie, used to store the release year of the movie
- **hasRating** – data property of Movie, used to store the average rating of the movie
- **biography** – data property of Person, used to store the biography of the respective person
- **height** – data property of Person, used to store the height of the respective person

Apart from the two datasets provided (*movies100.csv* and *names1000.json*), three additional ones are added to cover *Location*[1](*countries.json*), *Genre*[2](*genres.json*) and *Film Studio*[3](*production_companies.csv*) entities of our ontology. These datasets can be found in the *input/* path.

- From *movies100.csv* dataset, the following movie details are mapped to our ontology: **title** (as rdfs:label), **year** (as release year), **ratings**, **production_company** (as film studio), **director** (as movie director), **writer**, **actors**, **genre** and **country** (as location). The *rml* file to add these data can be found in the *rml/movies100.rml.ttl* path.
- From *names1000.json* dataset, the following person (actors, writers and directors) details are mapped to our ontology: **name** (as rdfs:label), **birth_name** (as name), **bio** and **height**. The *rml* file to add these data can be found in the *rml/names1000.rml.ttl* path.
- From *countries.json* dataset, the following location details are mapped to our ontology: **name** (as rdfs:label), **capital**, **currency** and **code**. The *rml* file to add these data can be found in the *rml/countries.rml.ttl* path.
- From *genres.json* dataset, the following genre details are mapped to our ontology: **name** (as rdfs:label). The *rml* file to add these data can be found in the *rml/genres.rml.ttl* path.

[1] World Countries data: https://www.back4app.com/database/back4app/list-of-all-continents-countries-cities/world-countries-dataset-api

[2] Genre list: https://www.back4app.com/database/paul-datasets/dataset-with-all-movies/movie-genres

[3] List of film production companies: https://en.wikipedia.org/wiki/List_of_film_production_companies

- From *production_companies.csv* dataset, the following film studio details are mapped to our ontology: **Company** (as rdfs:label), **Est.** (as established date) and **Country** (as location). The *rml* file to add these data can be found in the *rml/production_companies.rml.ttl* path.

For each one of the *rml* files, a separated *.ttl* file is created containing all the data imported from the datasets. In order to have all these data in one place, *.ttl* files are merged together with the initial (and extended) movie ontology. The final ontology is saved in the *output/movie.ttl* file and it is ready to be used for task 2.

Task 2: Enrich your Knowledge Graph through SPARQL

In order to further enrich our ontology created in task 1 of this project, the following 5 SPARQL queries are used:

1. For each person, set *rdf:type* to Actor/Writer/MovieDirector if he is actor/writer/movie director
2. For each actor, return movies in which the actor plays in
3. For each writer, return movies which the writer wrote
4. For each director, return movies which the director directed
5. For each film studio, return the total number of movies created
6. For each genre, return the total number of movies created

These queries can be found in the *input/query/* path.

Since all these SPARQL queries are CONSTRUCT queries, they add additional information to our ontology. The first SPARQL query will add to our ontology the information regarding if a person is an actor or writer or movie director, since this information is not initially provided in the *names1000.json* dataset. The second SPARQL query will add the information in which movies the respective actor plays in. Same goes for writers and directors where the information of which movies the respective writer/director wrote/directed. These first four queries will enrich further our ontology because having this information available for each person, makes it easier for us to create personalized content (profiles for each individual in our ontology).

Regarding the last two SPARQL query, the former will add the information of how many movies a particular film studio created and the latter is used to add the information of how many movies are of a particular genre. These queries add analytical information to our ontology which can be useful to report.

After all these queries are executed, the result sets are then mapped to the ontology from task 1. The ontology extended with the output of these SPARQL queries is then saved in the same *output/movie.ttl* file as in task 1.