## 2023Z Knowledge Graphs

# RML transformation of a static dataset into a knowledge graph

Illia Tesliuk, 296076

#### 1 Problem statement

This project aims to achieve the transformation of structured tabular data, where rows represent observations and columns represent attributes for these observations, into a knowledge graph represented by RDF subject-predicate-object triples. The goal of the transformation is to preserve relationships and semantics among different entities.

### 2 Project motivation

There are several motivations to convert static datasets into knowledge graphs. Firstly, contrary to tabular data, knowledge graphs allow for the representation of complex relationships between entities, thus enabling the discovery of connections that may not be apparent in tabular formats [1]. Secondly, with structured connections between entities, knowledge graphs make it easier to perform sophisticated search queries. Finally, knowledge graphs facilitate the integration of data from various sources and formats, offering a unified framework for the representation and organization of information. For example, each year a knowledge graph can be updated with the most streamed tracks within the last 12 months <sup>1</sup>.

### 3 Tools

One of the common ways to transform static data into RDF graph is to use RML (RDF Mapping Language). RML is a language specifically designed for expressing mappings between different data formats, such as CSV, XML, and RDF and RDF graphs. It provides a way to define how the data in these non-RDF formats should be mapped to RDF triples of subjects, predicates and objects.

However, RML mappings appear to have been designed with a focus on machine consumption rather than human comprehension. YARRRML [2] instead is a more common choice to create transformation mappings. YARRRML is a human-readable text-based representation for declarative Linked Data generation rules. A user defines mapping rules in a YAML file and a special parser converts them into an RML mapping which can be later used to create RDF triples.

I use yarrrml-parser<sup>2</sup> to convert YARRRML rules into RML format, and rmlmapper-java<sup>3</sup> to convert tabular data into knowledge graphs.

#### 4 Dataset

For this project, I've selected an open Spotify Songs database from Kaggle <sup>4</sup>. It consists of a CSV file with information about almost 30,000 tracks collected from the Spotify API. Each song is represented with 23 attributes, including basic information:

- Track name
- Album name

 $<sup>^{1} \</sup>verb|https://www.kaggle.com/datasets/nelgiriyewithana/top-spotify-songs-2023|$ 

<sup>&</sup>lt;sup>2</sup>https://github.com/RMLio/yarrrml-parser

<sup>3</sup>https://github.com/RMLio/rmlmapper-java

<sup>&</sup>lt;sup>4</sup>https://www.kaggle.com/datasets/joebeachcapital/30000-spotify-songs

- Artist name
- Playlist name
- Playlist genre
- Popularity
- Duration

Songs are additionally described by a variety of metrics that shape musical characteristics [3]. Spotify uses raw audio analysis to translate the track sound into a set of measurable metrics, that can be utilized in recommendation algorithms:

- Key overall key of the track
- Mode "minor" / "major"
- Tempo the overall "speed" of the track in beats per minute (BPM)
- Loudness overall track loudness in decibels (dB)
- Energy measures activity and intensity of the track
- Danceability describes how suitable a track is for dancing
- Speechiness detects the presence of spoken words in a track
- Acousticness describes the degree to which a song relies on acoustic instruments
- Liveness detects the presence of an audience in the recording (live / studio)
- Valence describes the musical positiveness conveyed by a track

However, this dataset requires some preprocessing before being mapped into RDF triples. Namely, some tracks involve multiple artists, which is usually indicated via 'feat' keyword in either track name or artist name. It would be difficult to create a custom parse function in YARRRML, therefore this preprocessing must be done in Python.

I've detected 5 types of entities to be extracted from the Spotify dataset, namely tracks, albums, artists, playlists and genres. For convenience of the YARRRML's mapping process, I've split the original CSV file into 5 separate tables to be sources of single entities.

## 5 Ontologies

I didn't manage to find a single ontology that would cover the majority of the track attributes presented in the dataset. I'm using Music Ontology Specification <sup>5</sup> (mo prefix) to describe such terms as Track, MusicArtist, Record, Genre and a duration feature. Additionally, I've selected the Playlist Ontology <sup>6</sup> (plo prefix) to describe the Playlist class. Since the floating-point metrics described above are unique to the Spotify dataset and are not present in any music-related ontologies, I've decided to create a custom Spotify ontology to cover these terms.

## 6 Methodology

To represent properties unavailable in mo and plo ontologies, I've defined a custom spo ontology in a turtle (.ttl) file with the https://mini.pw.edu.pl/kg/spo/# prefix. For each attribute present in the dataset, I've identified their domains and ranges and populated the spo ontology with the corresponding instances of rdf:Property (See Table 1). Numerical song attributes are defined to have a rdfs:Literal range, while those linking to other entity types are linked to the corresponding class.

<sup>5</sup>http://musicontology.com/specification/

 $<sup>^6 \</sup>verb|https://lov.linkeddata.es/dataset/lov/vocabs/plo|$ 

	Property name	Domain	Range
1	key, mode, energy, tempo, loudness, danceability, speechiness, liveness, acousticness, instrumentalness, valence, duration, popularity	mo:Track	rdfs:Literal
2	artist	mo:Track	mo:MusicArtist
3	album	mo:Track	mo:Record
4	playlist	mo:Track	plo:Playlist
5	${\it releaseDate}$	mo:Record	rdfs:Literal
5	genre	plo:Playlist	mo:Genre
6	subgenre	plo:Playlist	mo:Genre

Table 1: List of properties (rdf:Property) defined in a custom spo ontology.

I've used a single YAML file to declare YARRRML rules for every data source. The rules are grouped per entity and are assigned a unique key, namely *songs*, *artists*, *albums*, *playlists* and *genres*. Each entity section declares the name and format of its data source. YARRRML iterates over each row of the source file and applies rules defined in the corresponding section. Values of the columns are accessed using the following syntax: \$(column name).

IRI is used as the subject of the RDF triples for the entity. For example, the *song* prefix is defined as https://mini.pw.edu.pl/kg/Song/.Individual subjects are created by expanding the corresponding entity prefix with the ID read from the source file. For example, the subject rule for artists is defined as artist:\$(artistId), where *artist* is the entity prefix and *artistId* is the column name in the corresponding CSV file.

The goal of a predicate is to establish a relation between subject resources and object resources. Therefore, they are declared by providing the full property name that consists of the ontology prefix and the property's name within the ontology (e.g., foaf:name). Objects are declared with the help of data types (e.g., xsd:integer) and values that are read from the source file using the column access notation.

Alternatively, the value of the property can be read from a different source file. To achieve this, a mapping between the two entity types must be established. This is done using the mapping keyword followed by the name of the other entity. A new line starts with a condition keyword, which is followed by a declaration of a function in the next line.

I used the *equal* function, which takes two input parameters, each coming from a different source file. Each parameter is defined by a triple, including its name, the value read from the column of the source file, and an s or o letter, indicating whether the value of the column comes from the subject or the object of the triples. When the two column values are equal, a link between two entities is established. In total, I created rules for four mappings, namely (*Track,MusicArtist*), (*Track,Record*), (*Track,Playlist*) and (*Playlist,Genre*).

#### 7 Results

YARRRML rules described in the previous section were converted into RML rules and applied to the input dataset. The conversion resulted in a list of RDF triples written in a Turtle file. The resulting knowledge graph comprises 605,253 RDF triples. The output file appropriately defines entities and their properties, including mappings between different entities. To confirm the correctness of the graph, I ran 10 SPARQL queries on the knowledge graph. The queries, their descriptions, and outputs are presented in the Appendix.

#### 8 Conclusion

In this project, I transformed a static '30,000 Spotify Songs' dataset into an RDF knowledge graph. I defined YARRRML conversion rules in a single YAML file that establishes rules for every type of entity. These rules define how particular entities and their properties should be represented in the form of RDF subject-predicate-object triples. I created a custom ontology that defines properties

used in the Spotify recommender systems for representing the song's sonic characteristics. I tested the resulting knowledge graph with 10 different queries and ensured that mappings between different entities were established correctly. In conclusion, YARRRML is a convenient tool that allows the creation of user-friendly conversion tools.

### References

- [1] Katharina Brunner. Rdf, rml, yarrrml: A basic tutorial to create linked data from a relational database table, Mar 2024.
- [2] Pieter Heyvaert, Ben De Meester, and Anastasia Dimou. Generating linked data with yarrrml, 2018.
- [3] Nima Torabi. The inner workings of spotify's ai-powered music recommendations: How spotify shapes your playlist, Aug 2023.

### A SPARQL Queries

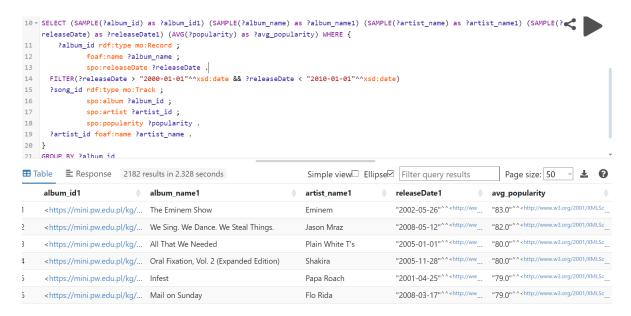


Figure 1: Albums released between 2000 and 2010, sorted by average popularity.



Figure 2: Ed Sheeran's albums, sorted by average popularity.

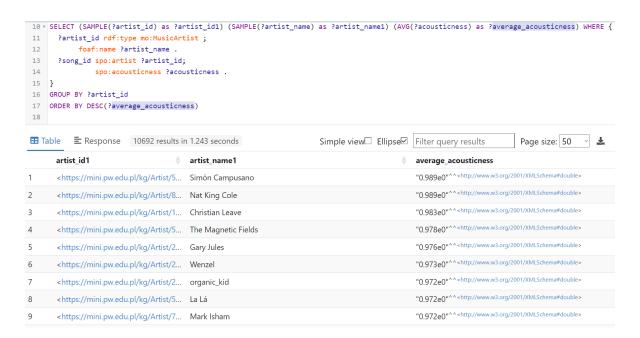


Figure 3: Artists, sorted by the average acousticness of their songs.

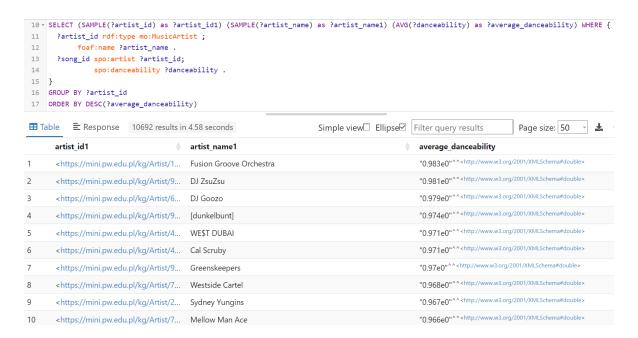


Figure 4: Artists, sorted by average danceability of their songs.

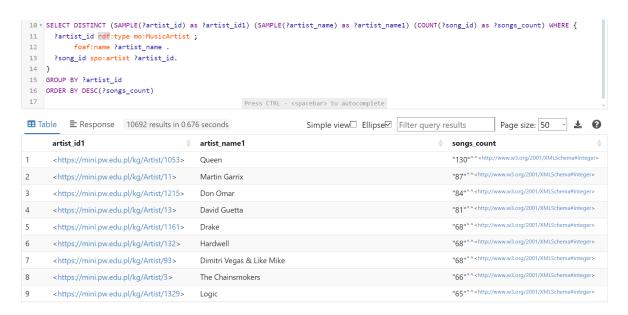


Figure 5: Artists, sorted by the number of songs they wrote.

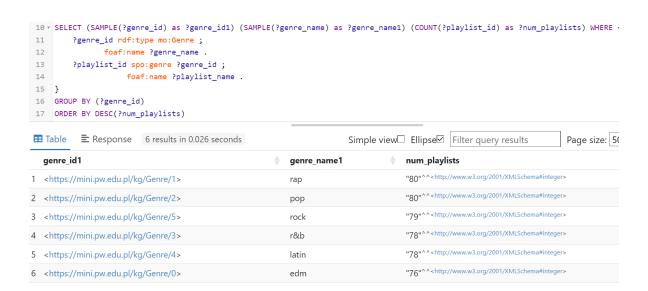


Figure 6: Genres, sorted by the number of their playlists.

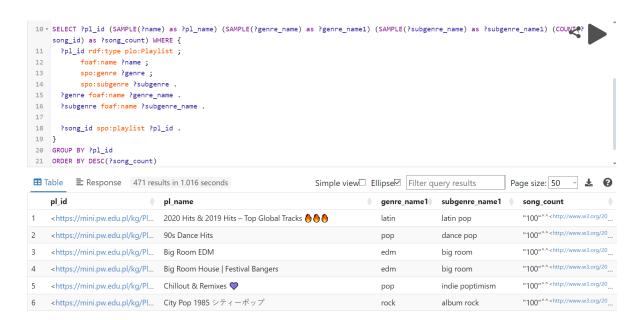


Figure 7: Playlists, sorted by their size (number of songs).

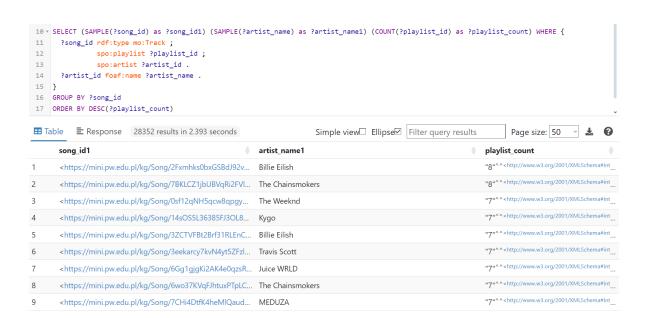


Figure 8: Songs, sorted by the number of playlists they belong to.

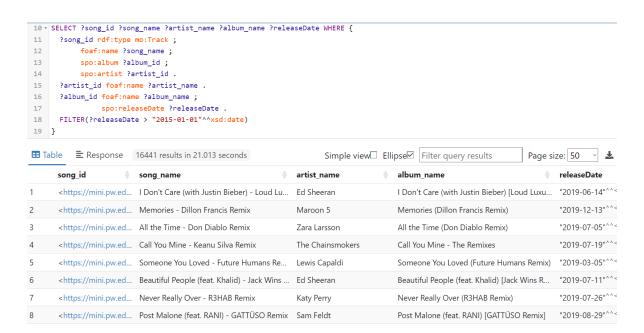


Figure 9: Songs released after 01/01/2015.

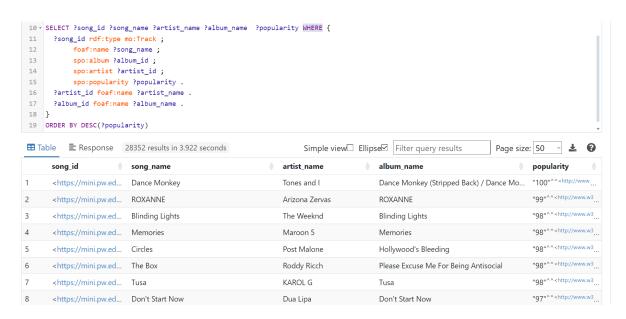


Figure 10: Songs, sorted by their popularity.