Background

A long standing and central challenge for musicians and artists and generally businesses online is to understand and model what users like or dislike about their art, and more particularly dive deep into the metadata of their production and extract interesting insights.We got inspired by the same problem-statement in the domain of Music. Our approach was to search for datasets online, that will help us create a knowledge graph for the Musical Domain and be able to answer questions and get insights.

We found MusicOSet, an open and enhanced musical dataset which offered rich features for us to do data engineering upon.

A brief description of the schema of the MusicOSet is provided below:

Artist Schema –

Followers Table: Information about the followers of the artist.
Popularity Table: The popularity of the artist. Values 0-100, 100 being most popular.
Artist Type: The type of the artists: one of "singer", "band", "duo" or "rapper".
Main Genre: The main genre in which the artist is associated with.
Genres: A list of the genres the artist is associated with.

Albums Schema –

Artists Table: The artists of the album.
Popularity Table: The popularity of the album. Values 0-100, 100 being most popular.
Tracks Table: The total number of tracks.
Album Type Table: The type of the album: one of "album", "single" or "compilation".

Songs Schema –

Artist Table: The artists who performed the track.
Popularity Table: The popularity of the track.
Explicit Table: Whether or not the track has explicit lyrics.
Song Type Table: The type of the song: "solo songs" or “collaborative songs”

Acoustic Features Table –

Pitch, Acousticness, Danceability, Energy, Intrumentalness, Liveness, Loudness etc are given as different properties of a song inside the table.

(To dig deeper into each acoustic feature and how its collected an introduction is provided here: https://marianaossilva.github.io/DSW2019/)

The total size of the dataset is 302MB, and contains 20,405 unique songs, 11,518 unique Artists and 26,225 unique albums spanning over 20 years!

We define our competency questions next, and define an ontology “music” which maps the uplifted dataset in a unique and meaningful way. The uplifted data is a subset of the total data.

Design
Query Descriptions

Q1. Which popular songs in Pop genre refers to "love" in their lyrics

Lyrics column is an object of Acoustic Feature connected through predicate :lryics. We filter from song popularity with literal "True" and map all these to song. Inside song we query through :hasFeatures and find the lyrics that CONTAINS "love".

Q2. What is the average duration of songs of type solo from the most popular artist of 2017?

The inner query finds the most popular artist for a particular year using a subject ArtistPopularity and mapping it to a subject Artist. The outer query finds all the songs connected to this artist, as each song has an album which has a release done by the artist. We take avg of the song duration and div it by 60,000 as it is in ms to convert it to minutes.

Q3. Which are the popular rap albums from an artist that have the higher average of "Speechiness" among its songs?

Using the transitive property :containsAcousticFeatures, we query the top 100 Albums on the basis of their speechiness.

Q4. Which band has the greatest number of popular albums in the last 30 years?

We count the number of albums that have an artist with artist_type as "band" and filter the album popularity to show results for values above 80. This result is grouped by the Artist subject, and counts the number of albums for that artist satisfying the above criteria.

Q5. Which is their most popular live song (liveness)?

This query builds on top of the previous query in Q4. It selects the first result from the above query and from the artist, maps all their songs, and the song's popularity and liveness and orders the result by decreasing liveness.

Q6. Who is the artist of the album which has the highest valued happiness in 2015 (valance)?

We map all artists to their releases and all releases to their respective albums, this gives us the album name w.r.t the artist name, now for all songs that have released in the year 2015, we get the valence of these songs and group them by their album name. The highest average valence is displayed.

Q7. How many songs released in 2018 that are danceable became popular?

All songs released in 2018, are filtered based on their dance values and mapped to song popularity having the Boolean (in this case Literal) value "True".

Q8. What is the average value of tempo of the 20 most popular electro genre songs?

For every artist having a genre "Electro", we query the mapping release and album and finally song connected to that. The result is finally ordered by decreasing value of the popularity metric of the song.

Q9. Which are the 5 most popular albums with high value of energy?

Using :containsAcousticFeatures transitive property, we find the albums, that have songs with energy greater than 0.5 . All such albums are grouped first based on album name and then on their average energies.

Q10. Who is the most popular artist that produced the higher number of sad songs (valence)?

We query the songs that have valence < 0.5, as that is a property that ranges from 0-1 indicating happiness of cheerfulness in a song. We map this to the songs, and count the number of such songs grouped by artists.