Course Project - Musician Knowledge Graph

Parii Dan
Department of Advanced Computing Sciences
Faculty of Science and Engineering
Maastricht University
Maastricht, The Netherlands

Abstract

Knowledge Graphs(KGs) have merged as pivotal components in facilitating a smooth, efficient experience in the realm of semantic technologies.KGs do so by organizing, linking, and extracting valuable insights from diverse data resources. This project aims to develop a KG that represents musicians, representing a diverse overview of the artist. Among others, the graph includes their lifetime sales, social media followings, and success on the charts. On the other hand, it also covers personal information such as their country of origin and number of children. The main focus of the project is the data retrieval aspect, various methods were combined to form a many-sided view of the artist. Methodologies such as querying WikiData, processing queried text with Large Language Models(LLMs), conducting sentiment analysis, and directly accessing datasets via Kaggle are explored. Afterward, the data is combined and a Knowledge Graph is created and validated. Finally, the graph was queried to highlight its utility as a tool for facilitating future research endeavors.

I. Introduction

Popular music influences and documents cultural, political, and economic contexts. The artists and companies behind the music substantially influence the world. Thus, the curation and documentation of our culture and its contributors is an important aspect of research, because understanding our past allows us to better understand ourselves and predict future trends. Historically, misdocumentation in music history is common; for instance, while many credit the likes of Elvis Presley with inventing Rock and Roll, it originated within African communities and was played in segregated areas, remaining largely unrecorded [1]. This way, its creators do not receive the accolades they deserve, and the masses are robbed of the truth. The goal of the project is the creation of a KG centered around artists, a KG that is suitable for both storing information and inference. In light of this, some concrete research questions will be enumerated, to better define our goals.

1) What characteristics of an artist should be added to provide a rich overview?
2) What is the evaluation methodology regarding a KG?
3) How can we evaluate the resulting Knowledge Graph as a tool for future research?

The project aims to integrate existing resources on music history and artist profiles into one comprehensive overview. The overview covers personal achievements such as Billboard Top 10 hits, but also more abstract characteristics such as the public image of an artist. By assembling this information, the project aims to serve

as a valuable resource for music researchers, offering information not only on the type of music but the background of the artists as well. Additionally, once populated, the schema will enable analysis to uncover historical trends, examine the demographics of music creators, and identify key factors driving their popularity and widespread influence.

Considering the task, a large part of the project focuses on the data retrieval aspect, including substantial pre-processing to allow for an integrated resource. Data was obtained from three different sources, namely: Kaggle, Wikidata, and Wikipedia. The Kaggle-sourced datasets are filtered to manage their extensive size. For the information queried from WikiData, classical NLP techniques are used to simplify features into a more suitable form for our focus. Furthermore, advanced techniques involving Large Language Models (LLMs) and transformer architectures are applied to Wikipedia's unstructured text, enabling the extraction of public sentiment regarding artists. Once the graph is integrated, quality assessments covering the Correctness, Conciseness, and Semantic Accuracy of the graph are performed. To evaluate the competency of the graph, some possible interesting insights are obtained by querying the final product.

II. Related Work

Previous works about Knowledge Graphs in music, have primarily focused on developing Recommender Systems [2]–[4], using the data structure to predict songs and enhance the listening experience. Our project takes a different approach, potentially offering an innovative application of Knowledge Graphs within the music sphere. As mentioned in section I, extensive documentation on music will inform our data collection post-schema creation. When it comes to documenting popular songs and artists, the Billboard 100 Chart [5] has been compiling a new chart every week since 1958. It provides an incredibly useful overview of what was popular at different points in recent history. As a data resource, it has been used in several capacities. Bayesian Models were used to predict list changes [6], using it to perfect and test ranking algorithms. Other works have also been more focused on predicting which songs will become number 1s [7], or even manufacturing a synthetic song that simply fits all the criteria [8]. Our project will use this list as a resource for determining what songs and artists were popular at the time, to populate our RDF Graph.

One of the most comprehensive sources of information on musicians can be accessed on user-generated platforms such as Wikipedia and WikiData. These resources are very important to the population of our graph, as they contain many specific details on artists, that could be obtained only by aggregating the work of users worldwide. While the information on these platforms may not be fully correct, this is a compromise we accept in exchange for access to such an extensive repository of data. The sheer volume and diversity of information available is what makes the validation impossible. There have been previous works that leverage these extensive resources for a more specific Knowledge graph population [9], [10].

III. Methodology

This project can be split into four distinct phases: 1) Schema creation, 2) Data Retrieval, 3) RDF Graph Creation, and 4) RDF Graph Validation.

A. Schema Creation

The development of the schema was guided by an assessment of available resources. An initial review was conducted on existing datasets from platforms such as Kaggle and Zenodo, where I identified some features for the graph. Following this, Wikidata's schema was evaluated for its utility as a data source, and its potential for interoperability. Wikidata was chosen due to its comprehensive coverage and interlinking with different domains. Lastly, the investigation was extended to cover abstract concepts from unstructured text.

B. Data Retrieval and Processing

The data was retrieved from three different sources: Kaggle, Wikidata, and Wikipedia. The data from Wikipedia and Wikidata was queried through their API services. From Kaggle, two datasets about an artist's lifetime sales and rankings on the charts were merged. There were challenges associated with the merging of the data, due to some artist's different spellings of the name. Additionally, the rankings dataset was reduced to contain only weekly top 10 ratings to reduce size. Data was then indirectly queried from Wikidata, namely quantitative features such as social media followings and number of children. Wikidata proved difficult to query automatically through their services, so a Python library called Pywikibot is used to automate work. To approximate an artist's public image, an LLM is used to choose interesting sections from an artist's Wikipedia page to reduce scope. Another model is used to classify the sentiment of each sentence from the text. For this purpose, out-of-the-box models are used from the HuggingFace library.

C. RDF Graph Creation

Regarding the creation of the graph itself, at first, the schema—including all classes, predicates, and their respective domains and ranges—was defined manually. Afterward, the GitHub Copilot plugin was used to generate most of the code regarding the inserts of the data as sentences in the graph, the generated code had to be tweaked considerably to achieve the desired result.

D. Evaluation

The graph is evaluated for its data quality, which is a multi-dimensional metric with the general concept of 'fitness for use'. The goal is to assess and ultimately improve the quality of the graph. The process covers checking different sources of possible issues, issues regarding reliability of source data, arising due to data integration, and incorrect use of vocabularies. More specifically, the quality dimensions evaluated in this paper are Accessibility, Intrinsic Quality(Syntactic Validity and Accuracy), Correctness, and Conciseness.

Regarding the Accessibility metric, an analysis was performed to check how many entities and predicates map to existing and popular frameworks. The Intrinsic metric of the graph was evaluated by checking for missing values and illegal numeric values, such as negative sales or number of children. To ensure Conciseness, we verified that each entity, such as labels and song titles, is described by just one label, eliminating any redundancy. Regarding the Correctness metric, our assessment focused primarily on how well the graph conforms to its predefined ontology.

These involved checks to ensure predicates were applied in their proper contexts, literals were accurately defined, and each entity was assigned a specific class, among other evaluative criteria.

Additionally, besides the quality of the graph, its competency will be evaluated, by creating queries to answer interesting questions. The questions are created to combine different characteristics of an artist which wouldn't be possible without the integrations done for this graph. The creation of the questions was facilitated by ChatGPT 3.5, with the ontology of graph provided in the context windows, and the questions asked in textual form.

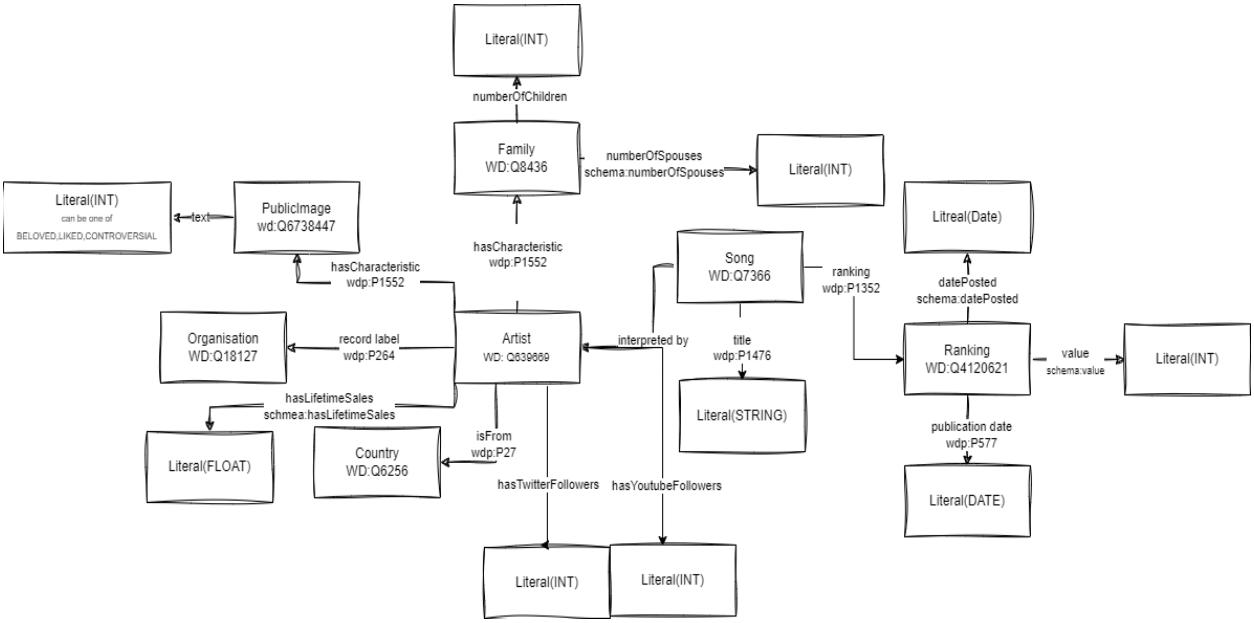

Fig. 1: Musician Knowledge Graph Schema

IV. Results

This section will describe the final version of the schema, and steps taken to raise the quality of the data, following this, interesting questions will be proposed and answered by querying the graph.

The schema was designed with simplicity, considering the time frame and scope of the project, but still captured the diverse characteristics of an artist (Figure 1). Characteristics such as the number of children and spouses are chosen to superficially account for an artist's personal life. Whilst, metrics such as public image, lifetime sales, and social media following account for their success as artists. Most of the entities and predicates of the graph were extracted from the ontologies of WikiData and Schema.org, with only a few predicates being too specific for the task. Namely, predicates 'schema:hasTwitterFollowers',' schema:hasYoutubeFollowers', and schema:hasLifeTimeSales' were created to describe the artist's characteristics in a simpler form, this kind of logic exists in WikiData, however, to replicate, quite a bit more entities and predicates are required.

The data retrieval and processing part is implemented in two Jupyter notebooks, namely *GraphConstruction* and *PublicImageAnalysis*. The process involved cleaning and merging data from 5 separate resources. 121 separate

Question	Answer(short)
Find how many artists has most followers on either Twitter/X or Youtube. Additionally find average number of followers on each platform	Youtube Followers: 44 Twitter Followers: 25 Avg Youtube Followers: 12.33M Avg Twitter Followers: 14.43M
Compare the number of sales an artist has, with the number of children they have , constraint the sales number in intervals of 50 million.	0-50 million Avg # of Children: 2.71 101-150 million Avg # of Children: 2.33 151-200 million Avg # of Children: 2.67 51-100 million Avg # of Children: 2.55 200+ million Avg # of Children: 2.43
Compare the public image of the artist with when their first song charted (see if nostalgia plays a factor). Order by Public Image Score.	Mariah Carey—PI Score: 111.69—Date: 1990-07-21 Frank Sinatra —PI Score: 80.54 —Date: 1966-06-04 Alicia Keys —PI Score: 74.83—Date: 2001-08-04 Shakira —PI Score: 65.51—Date: 2001-12-15 Elvis Presley —PI Score: 62.77—Date: 1958-08-04

TABLE I: Shows an example of questions determined to evaluate graph competency in a research setting.

artists were aggregated, all of whom had at least one top-10 song on the charts, proving their popularity.In the end, 5 pandas dataframes were developed, each having a common foreign key, that being the artists' name.

The creation of the graph is implemented in the notebook *GraphConstruction*, it has been performed in parallel with the quality assessment checks, which are implemented in the *GraphQualityAssesment* notebook. After code for a new characteristic to be inserted into the graph is written, the quality queries are run, then corrections are made to the construction implementation. This process was very useful, as the construction code had to be changed several times during this process to improve its validity.

To assess the competency of the graph, several interesting questions have been developed and answered by querying the graph. The queries and full responses can be found in the *QueryGraph* notebook. Three examples of this process have been included in Table I, only three were included for brevity's sake, more are implemented in the notebook.

The first question shows us that most musicians have more YouTube followers than Twitter followers, which is not surprising as YouTube is a much more popular platform. However, the average follower count is surprising, with more Twitter followers on average than YouTube. The second question compares the sales of an artist with the number of children. Here, we can see no deviations, with the number of children staying around 2.5 for all sales intervals. Lastly, the question regarding artists public image in comparison to their era shows us that the most beloved artists' first charting songs were quite some time ago, showing us a possible link that should be further explored.

V. Discussion

This project explored the creation and population of a KG, with its main focus being on the data retrieval aspect.

Obtaining different characteristics from different sources proved to be quite time-consuming. One of the issues faced in this process was the unreliable nature of the WikiData service, having very long wait times, just to end

up with errors, which is why WikiData was queried through a plugin as a substitution. Additionally, LLMs proved to be unreliable in extracting and formatting information from unstructured text properly

The evaluation methods of the graph proved to be very useful. Especially because the author is not very experienced with KG Creation, having concrete assessments of the graph proved instrumental in raising its quality.

The different areas covered by the graph are shown to help answer some interesting questions. For example, observing that artists with the best approximated public image had first charted more than 20 years ago, shows a possible link between nostalgia and public opinion. This topic, along with others proposed, can be studied in the future, using the KG with more advanced inference techniques such as Graph Neural Networks. Future work can focus not only on inference, but also extend the graph to contain other interesting features, or increase the number of artists covered.

VI. Conclusion

In summary, this project covered the creation of a Knowledge Graph, from data retrieval methods to assessing its quality and competency. The final product had passed all implemented quality checks and was able to answer several research-worthy questions. Moreover, the graph can be extended to include more advanced inference mechanisms and more data about the artists.

References

[1] B. Dreyer, "The african american legacy on rock and roll," *Waterloo Historical Review*, vol. 7, 2015.

[2] S. Oramas, V. C. Ostuni, T. Di Noia, X. Serra, and E. Di Sciascio, "Sound and Music Recommendation with Knowledge Graphs," *ACM Transactions on Intelligent Systems and Technology*, vol. 8, no. 2, pp. 1–21, 10 2016. [Online]. Available: https://doi.org/10.1145/2926718

[3] N. Bertram, J. Dunkel, and R. Hermoso, "I am all ears: Using open data and knowledge graph embeddings for music recommendations," *Expert Systems with Applications*, vol. 229, p. 120347, 2023.

[4] X. Liu, Z. Yang, and J. Cheng, "Music recommendation algorithms based on knowledge graph and multi-task feature learning," *Scientific Reports*, vol. 14, no. 1, p. 2055, 2024.

[5] Billboard, "Billboard." [Online]. Available: https://www.billboard.com/charts/hot-100/

[6] E. T. Bradlow and P. S. Fader, "A bayesian lifetime model for the "hot 100" billboard songs," *Journal of the American Statistical Association*, vol. 96, no. 454, pp. 368–381, 2001.

[7] C. Cibils, Z. Meza, and G. Ramel, "Predicting a song's path through the billboard hot 100'," 2015.

[8] M. DeAmon, "Predicting and composing a top ten billboard hot 100 single with descriptive analytics and classification," 2022.

[9] X. Zhang, X. Liu, X. Li, and D. Pan, "Mmkg: An approach to generate metallic materials knowledge graph based on dbpedia and wikipedia," *Computer Physics Communications*, vol. 211, pp. 98–112, 2017.

[10] A. Waagmeester, G. Stupp, S. Burgstaller-Muehlbacher, B. M. Good, M. Griffith, O. L. Griffith, K. Hanspers, H. Hermjakob, T. S. Hudson, K. Hybiske *et al.*, "Wikidata as a knowledge graph for the life sciences," *Elife*, vol. 9, p. e52614, 2020.