Final Report

- Building and Mining Knowledge Graphs -

Arthur Haas (i6241049)

April 5, 2021

Abstract

This project covers the conversion of the MusicBrainz database, a GDP per capita dataset and the world happiness index dataset into knowledge graphs to link them into one whole graph. Several queries were performed on this graph to answer research questions. During the project the main work lay in making the MusicBrainz data available by using docker, postgreSQL and RML mapper.

1 Significance

Music is a constant factor in the lives of many people. Artists create songs not only for pleasure but also to cope with emotional situations. The question that this project will try to answer is whether the amount of created music in a country can be explained by the population's happiness and country's economic growth.

Provided the amount of created music is influenced by above mentioned factors, then the opportunities of this research are seen in the creation of a new indicator of the population's stability in terms of happiness and economic growth. Not solving this research question deprives future decision makers from a new and easily measurable indicator. Besides these decision makers in governments and companies, also musicologists and humanities researchers(Oramas, Espinosa-Anke, Gómez, & Serra, 2018) are impacted by this research. The following list shows the research questions to answer achieve above significance:

- Main research question:
  - Can the amount of created music be used as an indicator for a country's happiness and economic growth?
- Further research questions to increase the analytical work:
  - How are the most hardworking artists in terms of number of songs?
  - Which country produced the most songs?
  - Find further complex but interesting queries, that can be executed on the musical dataset.

2 Related work

Olramas (et. al)(Oramas et al., 2018) identified the significance of applying data analysis techniques on musicological data. They were analyzing musical sentiment through data analysis techniques without taking into account the world happiness index, which will be added through this research project. Another added value is the comparison of the GDP per capita for every country.

3 Goal and specific objectives

The overarching goal of this project is two-fold. While a technical goal includes finding answers to the research questions and doing further research on the music dataset, the educational goal of this project is an increase of my understanding of knowledge graph techniques. The objectives pursued by this projects are:

- Download of all three datasets and making them usable (especially for the musicBrainz postgreSQL database)(Foundation, 2021).
- Conversion of relational datasets to knowledge graphs.
- Linking knowledge graphs.
- Querying resulting knowledge graph with SPARQL to answer questions.

4 Methodology

The listed objectives in 3 will be answered using three datasets: **MusicBrainz** is a large collection of music related data about artists, their work, etc.(Foundation, 2021). The **Happiness Report** gives an indication about the population's happiness((user on kaggle.com), 2020) and a report of the **GDP per capita** shows the economic status of a country(Bank, 2021). All of these datasets contain the country as a common entity, which will be used for linking.

For the conversion of above mentioned datasets into RDF knowledge graphs the *RDF Mapping Language (RMI)*[1] will be used. To achieve this, the programming language Python and thus the *r2rml-kit*[2] and *rdflib 5.0.0*[3] are considered. Information on SPARQL(Ali, Saleem, Yao, Hogan, & Ngomo, 2021) and from the lecture's textbook(Hogan et al., 2020) are used for further analysis.

Please refer to table 1 for the identified risks and how I plan on handling them.

Table 1: Risks

Risk	**Severity**	**Probability**	**Precaution & Avoidance**
As opposed to the prior research, above datasets are available as RDF graphs.	Severe	Low	Usage of those knowledge graphs. Instead of converting the datasets, further knowledge graph mining techniques from later in the course will be used.
MusicBrainz dataset too heavy to use on my machine.	Severe	Medium	Either choosing a lower detailed representation of data or using a remote virtual server for the data ETL steps.
Too many operation-system specific commands necessary	Medium	Medium (MusicBrianz postgreSQL)	Using a docker container to solve "works only on my machine" issue.

4.1 Setting up the project

Figure 1 gives an overview of the project setup. A larger version is provided in the appendix B. The setup will be elaborated on in this subsection.

[1] `https://rml.io/specs/rml/`
[2] `https://github.com/d2rq/r2rml-kit`
[3] `https://rdflib.readthedocs.io/en/stable/`

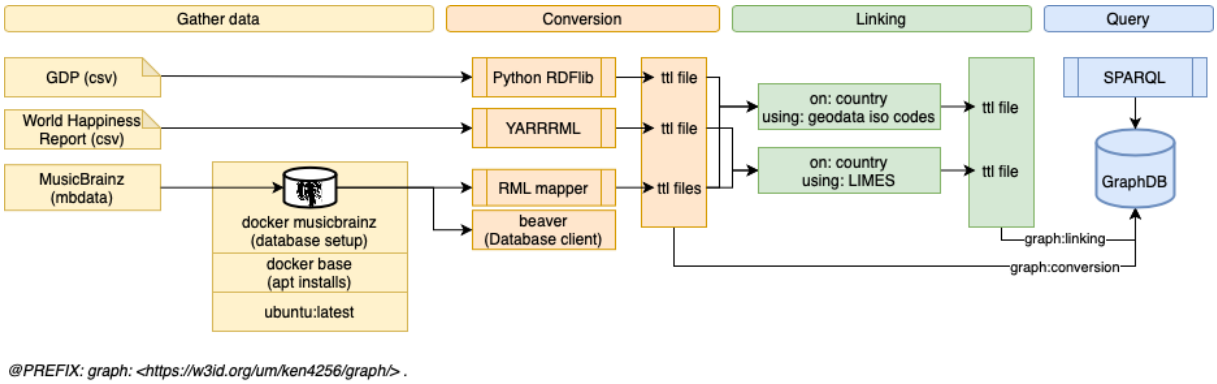

Figure 1: Project Setup

Download of the three datasets (MusicBrainz, Happiness Report, GDP Data) and making them usable were the first steps of the project. Downloading was straightforward. While the Happiness Report and GDP Data were combined around 0.5 MB, the size of the MusicBrainz dataset totals at around 50GBs. Hence, I had to make a choice of which part of the MusicBrainz data to use. I decided to use the core data for the knowledge graph and the statistics dataset as a plan B, because it is much smaller. Furthermore, to make the MusicBrainz dataset usable, a setup surrounding docker, postgresql and mbdata[4] was required. MusicBrainz recommended to use mbdata instead of their own MusicBrainz Server System, if the webserver is not requried. Since I need the dataset only, mbslave provided the postgres database setup only. To achieve this, I've setup an ubuntu docker base image and followed the tutorial on mbslave's GitHub repository. This was not complete in terms of commands to execute, hence I had to do a great amount of engineering research to make the database runnable. Details won't be mentioned here, because this report's goal is to show my knowledge graph skills. Instead, they can be found in the corresponding Dockerfiles and readme files.

4.2 World Happiness Report (YARRRML)

For converting the happiness report csv file into the turtle syntax, YARRRML was used. For the entity of *Country or region*, I used *schema:AdministrativeArea*, because items are not purely countries but areas with existing governmental body. The predicates were specifically defined using *owl:DatatypeProperty*. Due to incomplete function implementation and documentation, I have decided to perform the preprocessing step in python beforehand, i.e. creating an ID column which is URI conform for the subject. Meaning without spaces and bracket characters to avoid their unreadable encoding into % notation.

4.3 Data World.bank (Python rdflib)

For converting the GDP data, I have decided to experiment with a python library called rdflib. Using it was very flexible due to the programming language python. To convert the dataset, pandas was used to store the csv file and iterate it line by line. Most challenging part was the binding of a namespace to make it appear as a correctly writting prefix in turtle.

4.4 MusicBrainz (RML mapper)

The most challenging dataset was MusicBrainz due to its complexity with over 200 tables. I have decided to use plain RML. To create a working connection to a postgresql server I had to look into RML test cases, because the documentation is too sparse[5]. In addition, I decided to

[4] `https://github.com/lalinsky/mbdata`
[5] `https://rml.io/test-cases/#rmltc0001a-postgresql`

split the RML mapping into files corresponding to certain entities like country, releases, release groups and artists.

5 Milestones & deliverables

Milestones - Starting this week, around five weeks of time are available:

1. Week 1
    - Submission of project proposal
    - Download datasets.
    - Focus on group assignment 1.
2. Week 2
    - Review feedback for project proposal.
    - Make datasets usable within postgreSQL.
    - Convert all three datasets into knowledge graphs.
3. Week 3
    - Link three knowledge graphs.
    - Do a basic exploration to get used to the knowledge graph and to find eventual errors.
    - Focus on group assignment 2.
4. Week 4
    - Perform SPARQL queries to answer questions.
    - Perform SPARQL queries to find further insights.
5. Week 5
    - Write final project report.
    - Prepare all deliverables for submission.

Deliverables:

- Converted and linked knowledge graph.
- Local database postgreSQL with used data.
- Project report
- (optionally) docker image configuration file

6 Anticipated results

Main outcomes of this project will be a new knowledge graph connecting three datasets and thus resulting in new knowledge for the music industry. From that it is expected to find an answer to the question, whether or not the amount of created music is a good indicator for the country's happiness and economic growth.

7 Results

Most figures of this section are moved into the appendix. Figure 2 shows the initial thought process about how to build the final knowledge graph. It was mainly based on the three input datasets. An entity-relationship graph for the MusicBrainz database can be found on their website.[6]. To allow for more explainability, a reificaiton technique was used for the GDP data, as illustrated in figure 3.

After conversion and linking, the data was loaded into GraphDB. Figure 4 shows an overview of some entities and relations. Several queries were performed to answer the research questions. To answer the main question, also graphs were generated, which can be found in figures 5 and 6. All SPARQL queries can be found in the deliverables at location *sparqlSPARQL_queriesŕq*.

Results of these queries are listed now:

Query 1 The amount of released music in the year 2019 does neither correlate with the countries GDP per capita nor with their happiness score. Thus, the main research question can be answered with no. Countries with most releases in the year 2019 do not have the highest GDP or happiness score value.

Query 2 Artists named *Grant MacDonald*, *sonnov*, *Zippy Kid* and *Ludwig van Beethoven* have produced the most release groups in the year 2019, which is equivalent of saying, they have published the most individual releases. Detailed look into the data reveiles, that these artists create "noise" melodies[7] on a regular basis.

Query 3 To answer another research question, a similar query to *Query 1* can be executed. It shows, that the most releases were published in the US, Japan and the United Kingdom.

Query 4 Most releases were published by real male and female people, with around 40.000 releases and 10.000 respectively. Afterwards with around 300 releases a fictional female character produces releases.

Query 5 Most common languages for publishing releases are English, Japanese and Spanish. Some releases are tagged with *Multiple languages* indicating, that the release was not limited to one language only.

Query 6 Analysing the most international artists in terms of number of countries, they publish their releases, shows familiar names. David Guetta, Grimes and Hana are most international with 225 countries each.

8 Discussion

Making the MusicBrainz database usable took very much effort, because the provided documentation by MusicBrainz was either outdated or did not fit to my usecase of database only. They provide a MusicBrainz Server as well, but running this on my machine would not be possible. Hence, I've decided to use the MusicBrainz Database Tools, called mbdata[8]. Their documentation was outdated and incomplete, resulting in extra research and many trials. Finally, deploying the core data from MusicBrainz in a postgresql server within a docker container worked out very well.

[6] `https://musicbrainz.org/doc/MusicBrainz_Database/Schema`
[7] `https://musicbrainz.org/artist/5fde3528-8641-4578-a031-37748a7c20ca`
[8] `https://github.com/lalinsky/mbdata`

As figure 2 shows, for as many attributes as possible an IRIs was used to greatly increase connectivity in the resulting knowledge graph. This increased explainability, because those IRIs include labels and comments. For example the IRI for language is based on an ISO-3 code but contains a human readable label. This also increased the speed of SPARQL query executions, because GraphDB could rely on IRIs as opposed to strings, which would have to be compared across all entities. On the other side, creating IRIs instead of simple Literals was much more time consuming.

Linking all three datasets was an interesting task. For the link between GDP and MusicBrianz a mapping file from geonames was used[9], because the link was made between and iso-2 and an iso-3 country code, which are static. For the link between the happiness report and MusicBrainz on the other hand, LIMES was used. It successfully found a link for 150 countries. The remaining 6 were written very differently or were not valid countries, e.g. the countries *Democratic Republic of the Congo* and *Congo* were stored as congo_(kinshasa) and congo_(brazzaville). Also the area of "palestinian_territories" is a commonly referred to area, but not a country to link an iso-code to.

To approach the main research question the first query provided the necessary dataset and a correlation analysis answered it. Figure 5 shows three line graphs for every country in descending order by the number of releases in year 2019. Subfigure 1 shows the actual number of releases, subfigure 2 the Happiness Score and subfigure 3 the GDP per capiata. While this figure shows a somewhat similar behaviour for all three measures, the real indicator to answer this question in displayed in figure 6. The correlation analysis shows a correlation around 0.5 for both, happiness score to release count and gdp to release count. Hence, the number of releases is of certain similarity to the other two measures, but it is not large enough to explain it.

9 Conclusions

In conclusion it can be safely said, that the amount of published music in a country does not influence the countries GDP per capita value nor its Happiness Score. But on the other hand, a number of insights were taken from this project, with the main one being, that the MusicBrainz database provided a great practise example to convert a dataset into a knowledge graph. In addition, I've experimentet with three different kinds of conversion techniques, being YARRRML, RML mapper and the python library RDFlib. All of these were helpful in reaching my goal.

[9] `https://download.geonames.org/export/dump/countryInfo.txt`

References

Ali, W., Saleem, M., Yao, B., Hogan, A., & Ngomo, A.-C. N. (2021). A survey of rdf stores & sparql engines for querying knowledge graphs. *arXiv preprint arXiv:2102.13027*.

Bank, T. W. (2021). *Gdp per capita (current us$) [csv file].* (Available from `https://data.worldbank.org/indicator/NY.GDP.PCAP.CD?end=2019&most_recent_year_desc=false&start=2010&view=chart`)

Foundation, M. (2021). *Musicbrainz [postgresql database].* (Available from `https://musicbrainz.org/doc/MusicBrainz_Database`)

Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., de Melo, G., Gutierrez, C., ... others (2020). Knowledge graphs. *arXiv preprint arXiv:2003.02320*.

Oramas, S., Espinosa-Anke, L., Gómez, F., & Serra, X. (2018). Natural language processing for music knowledge discovery. *Journal of New Music Research*, *47*(4), 365–382.

(user on kaggle.com), M. A. (2020). *World happiness report up to 2020 [csv files].* (Available from `https://www.kaggle.com/mathurinache/world-happiness-report`)

A APPENDIX: Graphs

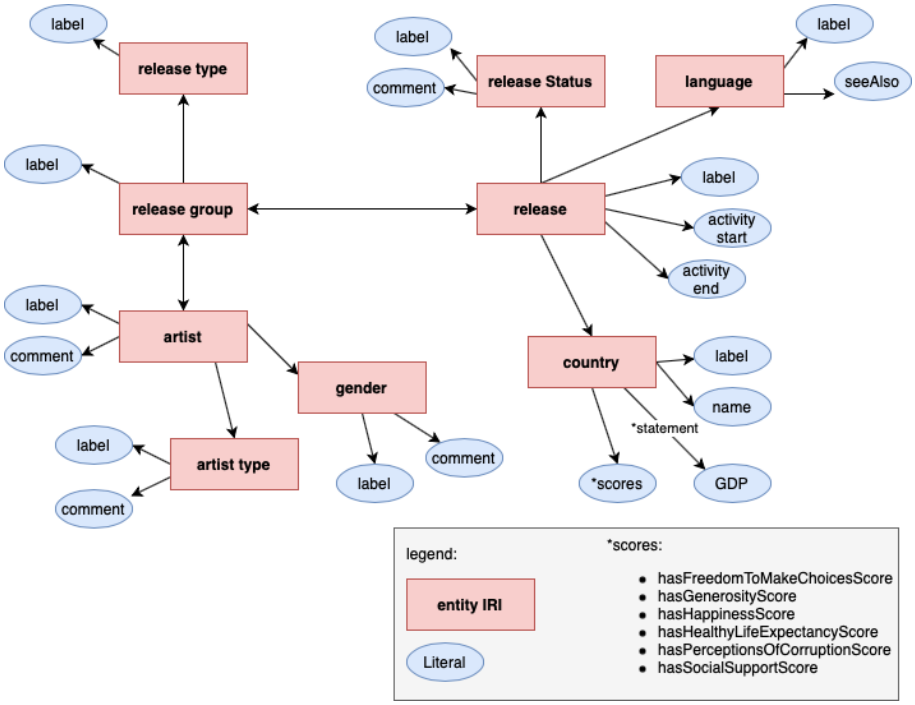

Figure 2: Knowledge Graph Conceptionalisation

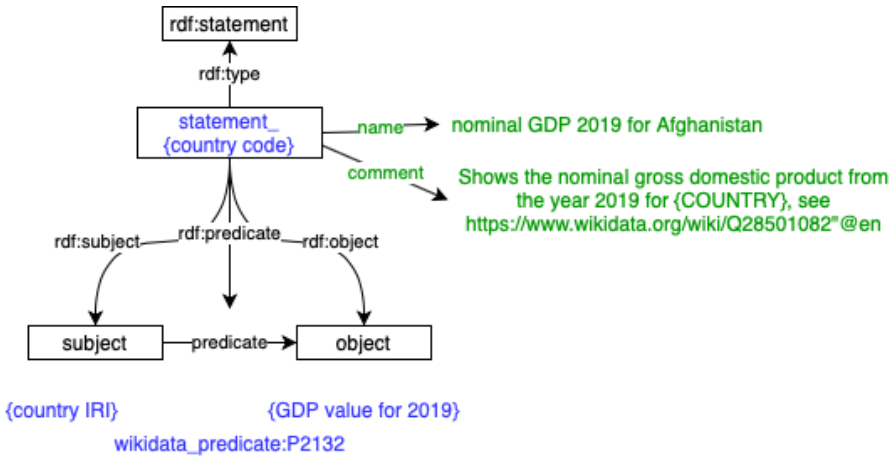

Figure 3: Reification Statement

Figure 4: Example Graph from GraphDB

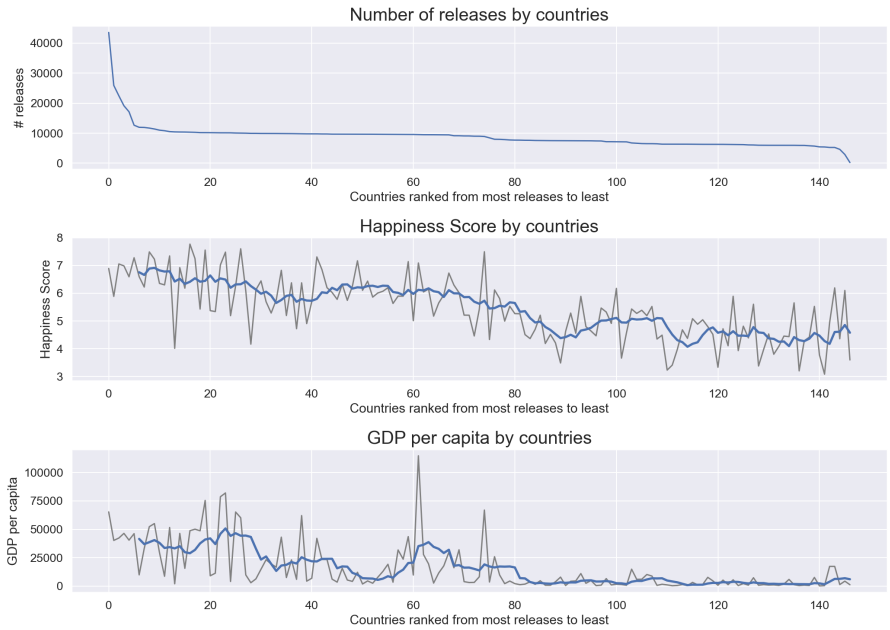

Figure 5: Query 1

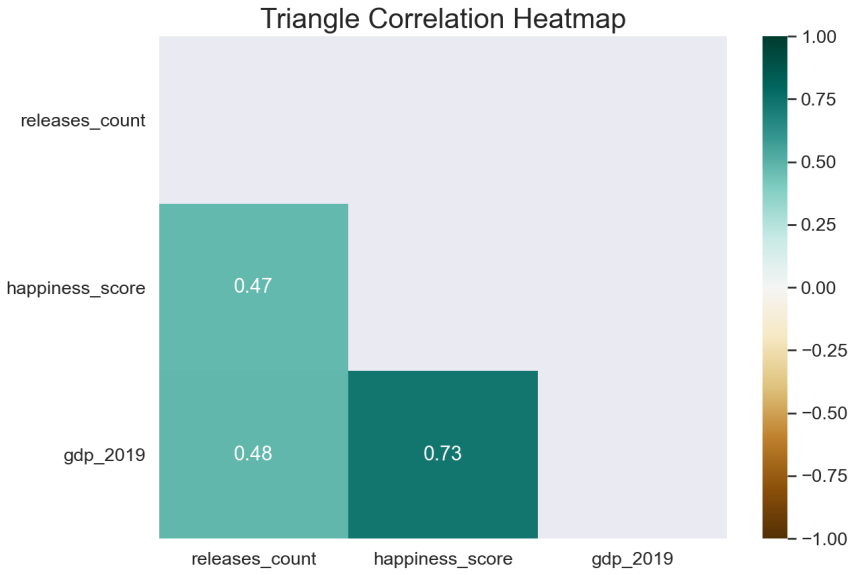

Figure 6: Query 1 correlation

B APPENDIX: Large Project setup

Larger version of the figure used in section 4.1.

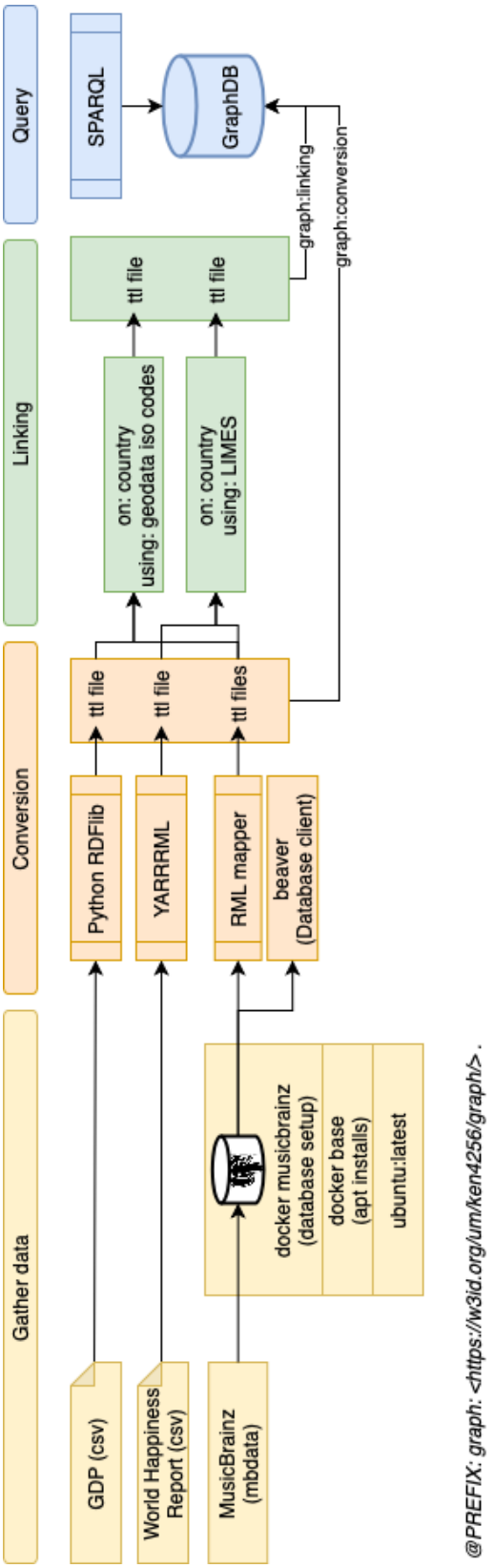

Figure 7: Project Setup