BMKG Individual Project

subtitle

Emese Szakály (i6259862)

April 5, 2021

1 Introduction

This project is about Marvel movies and their superheroes. During my work I will mainly be using knowledge gained from the lectures and labs, and the recommended textbook [4].

2 Significance

The task I would like to tackle is creating a knowledge graph from existing datasets about Marvel movies and superheroes and then using this new form of knowledge to find some interesting facts. The significance of this problem lies in the culture of superheroes. This magical world plays a big role in many people's life, since these superheroes are not just colorfully dressed characters who fight in order to defeat the devil, but they symbolize different parts of our civilization that we would desire to be real, for example truth and justice. By gaining new information about the nature of these superheroes and the success or interest of the movies they play in, we can learn about what most of the people would like to see, e.g. what characteristics the most popular heroes have could help to create new superheroes that are favoured by the majority.

3 Related work

I found several datasets about the Marvel movies, for example on `https://www.kaggle.com/`, but no RDF made from them, therefore I think it would be useful for others to have a knowledge graph with united information from multiple datasets, where they can run queries about what they are interested in. There are related papers about graphs created from the superhero network based on the movies they play in [2] or [1], also about creative visualization of these networks and some statistics [7], but not really about having information from both the movies and characters and connect them.

4 Goal and specific objectives

After linking the datasets I will be working with, I would like to come up with interesting questions and find answers to them by querying in the new, linked dataset. So far I came up with questions such as are there surprising facts regarding the ratings / profit of a movie and the budget they spent on it? Or after creating a social network of the heroes based on [6], what information can we find from those characters that appear in more movies? What is the most common hair color of the most popular heroes? Are they rather good or bad characters? Which are the most popular superpowers used in the comics that are the most popular?

5 Methodology

First, I am planning to build up an RDF model from each of the three datasets ([3], [5], [6]), or the relevant files from them I choose. For this I would use the RDF Mapping Language (RML) [1], and as a tool for the mapping the RML processor [2]. I thought of some ideas for this already, such as hero A playsWidth hero B, or hero playsIn movie. After this I would like to link the datasets, with the help of LIMES [3]. When all of this is done, using SPARQL [4], I plan to answer interesting research questions based on the newly created knowledge graph. About possible risks that can occur during the project, see Table 1.

Possible risk	Solution
One of the datasets already exists in RDF format, but I could not find it, or it gets created by someone during my work.	Considering, that I can still do the transformation to RDF on the rest, it is not a huge problem.
The dataset containing the ratings of Marvel movies [3] cannot be fully integrated with the dataset containing the superhero descriptions [5].	I will try to find as much overlap as possible, but if it still does not seem enough, I will try to search more to find a more fitting dataset having the ratings of these comics.

Table 1: Risk analysis

6 Milestones and Deliverables

During the first week, I would like to finish with the conversion of the datasets to RDF. By the end of the second week, I intend to finish with linking the datasets, and the upcoming week I plan to answer my research questions. Last week I will write the report about my results. This way, even if unexpected factors appear, my work should still be done before the deadline.

7 Anticipated results

Probably the hardest task will be combining the eight files belonging to the superhero dataset [5], because the information is spread among those and I will have to find an optimal way to unite them. Nevertheless, I expect to find some interesting facts about the Marvel Universe, relationship of the superheroes, patterns in the characteristics of these heroes, and maybe surprising facts about the popularity of the comics and the used budget for it.

8 Results

As Figure 1 shows, I choose 4 .csv files from the aforementioned 3 datasets, and created one .json file as input as it deemed necessary. After mapping these data files with 3 different methods to RDF, I did linking on two pairs of datasets, and then on the linked knowledge graphs I ran SPARQL queries. The program files are attached in the .zip file, and the names of files on Figure 1 help with navigating among the files.

[h]

[1] https://rml.io/specs/rml/

[2] https://github.com/RMLio/rmlmapper-java

[3] https://aksw.org/Projects/LIMES.html

[4] https://www.w3.org/TR/sparql11-query/#idp2427544

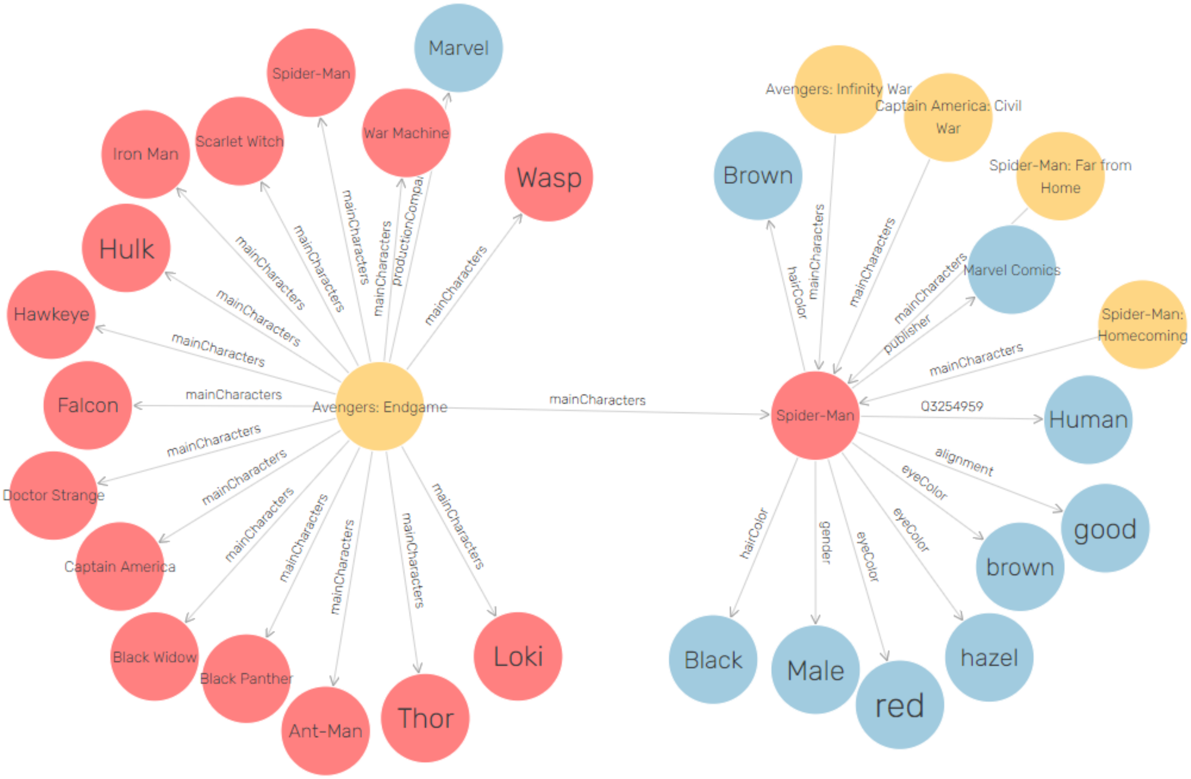

Figure 1: Example from GraphDB showing different type of entities with their properties and links.

9 Discussion

There were several challenges to be faced during this project. For example, the character names in the hero-network and in the other files about characters were totally different, therefore linking with LIMES did not work among them. This is why there are two graphs as final result, one about the hero-network, and the other regarding the movies, statistics about them, superheroes and information about them, and link between movies and their main characters. For the latter I used LIMES two times, for the files where it was possible to get reasonable results (where the character labels and the heros who appeared were similar). Linking gave 35 matchings for using LIMES for $output_marvel_characters.ttl$ and $output_character_label.ttl$, and 582 matching between the character labels of $output_marvel_characters.ttl$ and $output_characters.ttl$. This was a satisfying result, but for example for the linking of the character statistics file and the hero network, did not work at all, no useable result was derived.

One challenge was that there was no available dataset about listing the Marvel movies and all their characters. This way first I could not connect the movies and the characters, but then I created a.json file manually to link the main characters of the movies to the movies. I wanted to do the mapping to RDF with YARRRML, but since I could not find documentation about how to split the items of the output, I had to use Python to finish up this mapping. After this I could link together all datasets except for the hero-network.

I wanted to learn the most from this project, therefore I used three different methods to do the RDF mapping. Two of them were from the class (RML and YARRRML), but I used a new one, too, called RDFLIB[5], which is to be used in Python.

[h]

GraphDB helped a lot in realising mistakes in my mapping or linking, and also, of course, to help me see through the project easier. Figure 2 shows an example of how the entities are linked and what properties they can have.

For SPARQL queries, I tried to use exciting questions, which are challenging, as well as less likely to be found

[5] https://rdflib.readthedocs.io/en/stable/index.html-rdflib5.0.0documentation

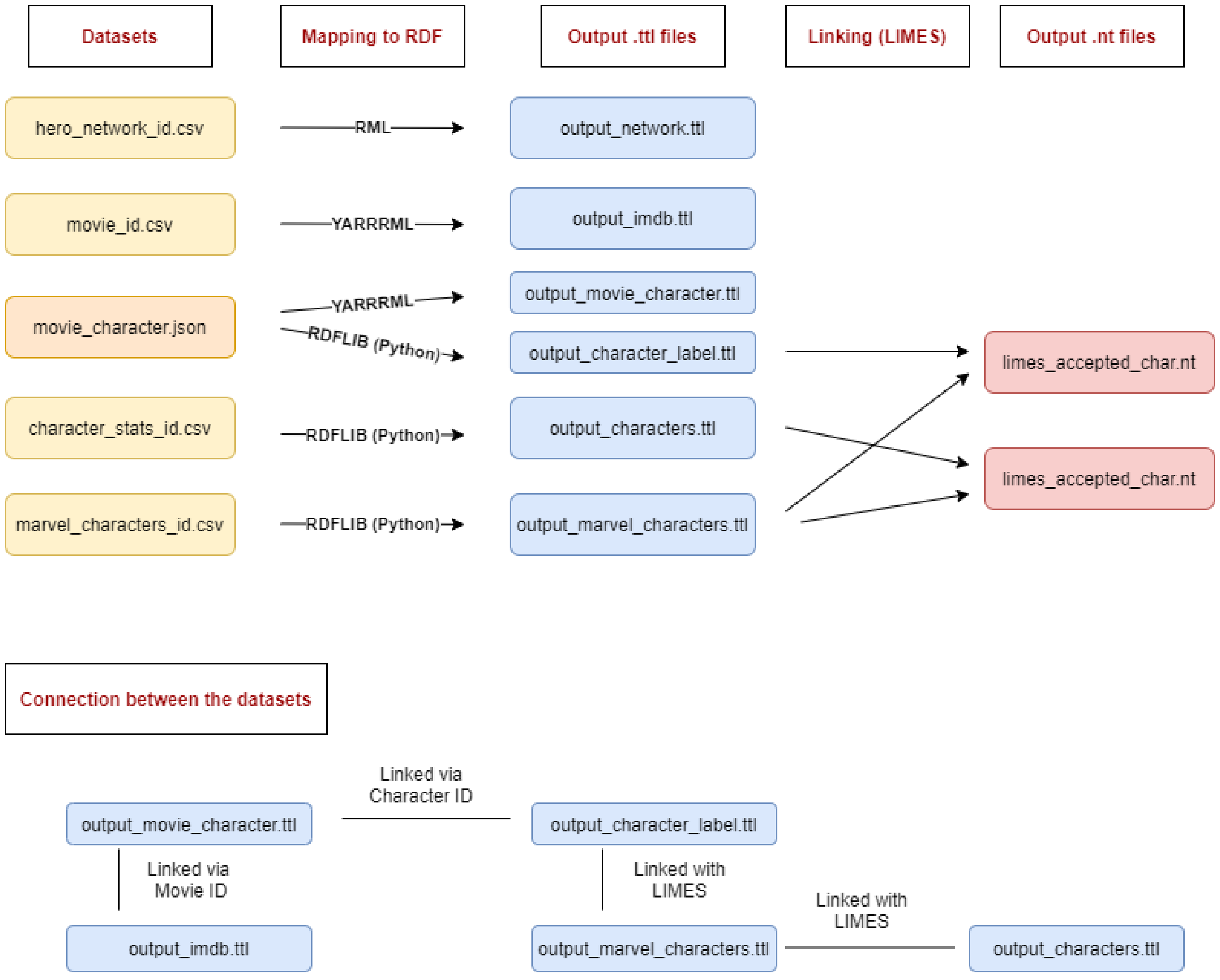

Figure 2: Process of RDF conversion and linking of the datasets.[6]

elsewhere. I used a tutorial (`https://www.stardog.com/tutorials/sparql/`) which helped me a lot in writing the queries. The queries are:

- Which 10 superheros have the most connections in the hero-network?
- Which heroes are the connections of Captain America (the one with the most connections)?
- Which 6 movies have the highest gross Worldwide?
- Which heros play in the movie which has the highest gross Worldwide?
- What is the alignment of the heros that play in the movie that has the highest gross Worlwide?

The answers to these queries can be found in the attachment.

A great idea to continue this work would be to link the hero-network with the other graph.

10 Conclusions

Overall, I believe this project reached this aim, I used what I learnt on the course, also more, with RDFLIB, for example. I managed to build RDF from different types of files, to apply linking and run SPARQL queries on the knowledge graph, as planned. I think that the answers found to the queries are interesting and partly new, and I hope that others will be able to run queries on the graph according to their taste in the future, or to continue this work with extension.

References

[1] Marvel cinematic universe.

[2] Cory Everington. Marvel cinema universe network analysis.

[3] Leonardo Henrique. Marvel vs dc, 2019. Available from kaggle: `https://www.kaggle.com/leonardopena/marvel-vs-dc`.

[4] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutierrez, José Emilio Labra Gayo, Sabrina Kirrane, Sebastian Neumaier, Axel Polleres, et al. Knowledge graphs. *arXiv preprint arXiv:2003.02320*, 2020.

[5] Danniel R. Marvel superheroes, 2018. Available from kaggle: `https://www.kaggle.com/dannielr/marvel-superheroes`.

[6] Claudio Sanhueza. The marvel universe social network, 2017. Available from kaggle: `https://www.kaggle.com/csanhueza/the-marvel-universe-social-network`.

[7] Amanda Chen Xingya Wang and pliang. Marvel universe visualization.