STREAMFUSION RDF

Presented By : Rodrigo Araya, Patricio Espinoza and Matías Rivera

Group 4

DCC UChile | 2024

Introduction

Overview

- Selection of datasets
- Preprocessing of datasets
- CSV to RDF
- Query creation

Objective

Users now have access to numerous streaming platforms like Netflix, HBO Max, Amazon Prime Video, and more. However, the availability of movies and series varies by region and platform, making it challenging to find consolidated information. This project aims to address this issue by integrating and analyzing data from different streaming services.

Preprocessing

- Select columns of interest.
- Remove missing values.
- Add the platform to which the data corresponds.

CSV to RDF

- We tried using Tarql, but we prefer the use of RML.
  - We defined rules to convert the data from CSV format to RDF triples.

Difficulty

Generating triples from columns that contained more than one entity.

Queries

- Our RDF data was analyzed through a series of SPARQL queries executed using Apache Jena Fuseki.

The following queries were performed to extract meaningful insights from our RDF dataset:

1. Names and count of platforms hosting each movie

2. Movies available on the same platform with country-specific variations

3.Predominant genre on each platform

4. Popular genres in Chile

Queries

- More queries:

5. Number of movies available by country.

6. Count of movies shared across platforms.

7. Number of movies released per decade.

Conclusion

- In this project we applied semantic technologies like RDF and SPARQL to integrate and analyze data from streaming platforms.
- In the future, we could incorporate additional datasets from other platforms to build a more robust query system. Additionally, we could develop an endpoint to allow users to perform customized queries.

THANK YOU

Presented By : Rodrigo Araya, Patricio Espinoza y Matías Rivera