Gustavo Adrien Polli

Aluno de Tecnologia em Análise e Desenvolvimento de Sistemas - Turma 2018
@ Faculdade de Tecnologia da Unicamp (FT)

M&SBox (Movie & Soundtrack Box)

21 de novembro de 2019

Visão geral

A obtenção de dados a partir de fontes em HTML muitas vezes é uma tarefa penosa que requer um grande esforço por parte dos programadores. Falta de padrão nas páginas e nos dados são alguns dos maiores incômodos para nós.

Durante a obtenção de dados para a elaboração deste projeto, esbarrei exatamente nesses problemas. A solução encontrada foi obtê-los manualmente (costumamos chamar esse ato de “tirar na unha”), um a um, em número reduzido, em vez de fazer scraping ou crawling por intermédio de código.

Objetivos

1. **Pesquisa de Faixas:** o objetivo principal, e mais importante, desse projeto é a pesquisa de faixas a partir dos dados dos filmes.
2. **Pesquisa de Álbuns:** a obtenção dos nomes dos álbuns é consequência direta - e não menos importante - da pesquisa de faixas a partir dos dados dos filmes.
3. **Outras pesquisas possíveis:** pesquisa de compositores a partir dos dados dos filmes, pesquisa de filmes, pesquisa de atores e pesquisa de diretores.

Especificações

Para o documento RDF Turtle, foram utilizados os seguintes dicionários RDF:

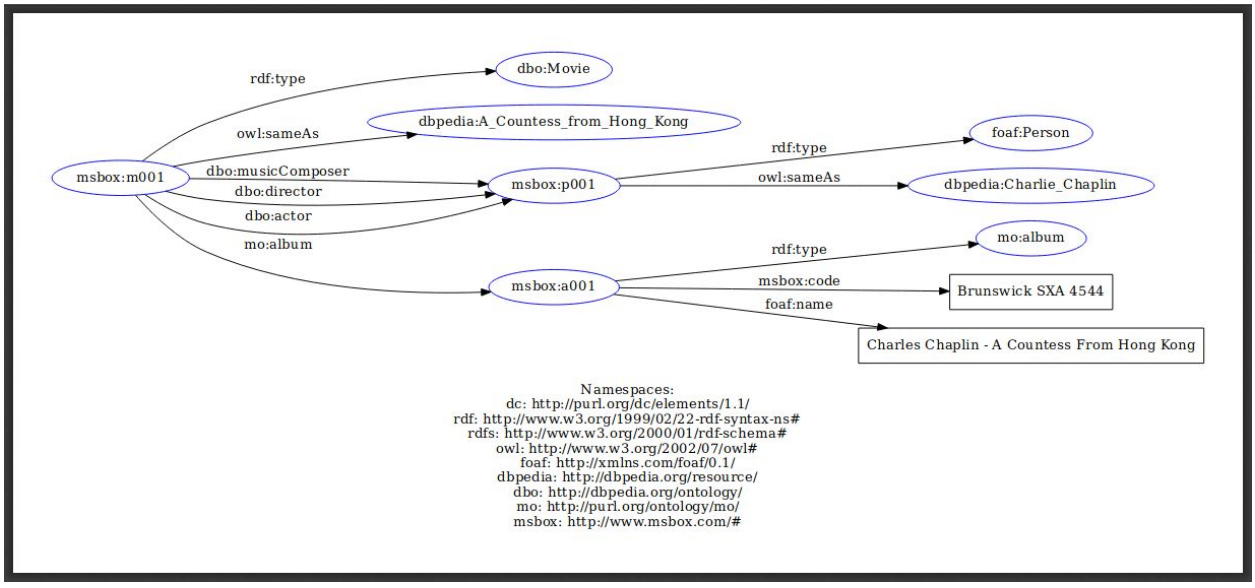
1. **Dublin Core (dc):** vocabulário que permite a descrição de metadados (dados sobre dados).
2. **RDF Syntax (rdf):** vocabulário que define os termos necessários para que sejam criados vocabulários em RDF.
3. **RDF Schema (rdfs):** vocabulário que descreve classes e relações entre elas.
4. **Web Ontology Language (owl):** vocabulário que permite a descrição de ontologias (representações de fatos e regras sobre um domínio de conhecimento).
5. **Friend of a Friend (foaf):** vocabulário que define termos para descrever pessoas, suas atividades e seus relacionamentos com outras pessoas e objetos.
6. **DBpedia Resource (dbpedia):** vocabulário que representa os dados contidos na Wikipedia em formato RDF.
7. **DBpedia Ontology (dbo):** vocabulário que define as ontologias utilizadas na DBpedia.
8. **Music Ontology (mo):** vocabulário de ontologias para criação de dados relacionados à música e afins.
9. **MSBox (msbox):** vocabulário criado para este projeto. Seu objetivo é permitir a manipulação de dados sobre filmes, trilhas sonoras, faixas, etc.

Termos criados

Obs.: “---” é um índice iniciado em 001 e incrementado a cada novo dado obtido.

1. **msbox:m---:** ontologia criada para descrever dados sobre filmes.
2. **msbox:p---:** ontologia criada para descrever dados sobre pessoas, sejam essas relacionadas a filmes ou trilhas sonoras.
3. **msbox:a---:** ontologia criada para descrever dados sobre álbuns.
4. **msbox:code:** dado literal usado para descrever o código do álbum.

Modelo de dados (reduzido)



O RDF final possui 368 triplas.

Exemplos de consulta SPARQL

1. **soundtrackQuery.py:** pesquisa todas as faixas relacionadas aos filmes.
2. **movieByTrackQuery.py:** a partir de uma faixa informada pelo usuário, o script pesquisa dados sobre filmes e compositores.

Dependências

1. **rdflib:** biblioteca que permite trabalhar com RDF.

Repositório

https://github.com/gapolli/MSBox

Fontes de Dados

1. Filmografia de Charlie Chaplin

   http://www.adorocinema.com

2. Dados sobre soundtracks

   http://www.soundtrackcollector.com

3. Dados abertos conectados

   http://dbpedia.org (DBPedia)

"Não há conhecimento que não seja poder". **(EMERSON, R. W.)**