INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES LYON

DBM1: Harry Potter Movies

IST – Database and Data Mining

1st Semester 2023/2024

Apopei Cristina

Popa Alexandra

I. Introduction

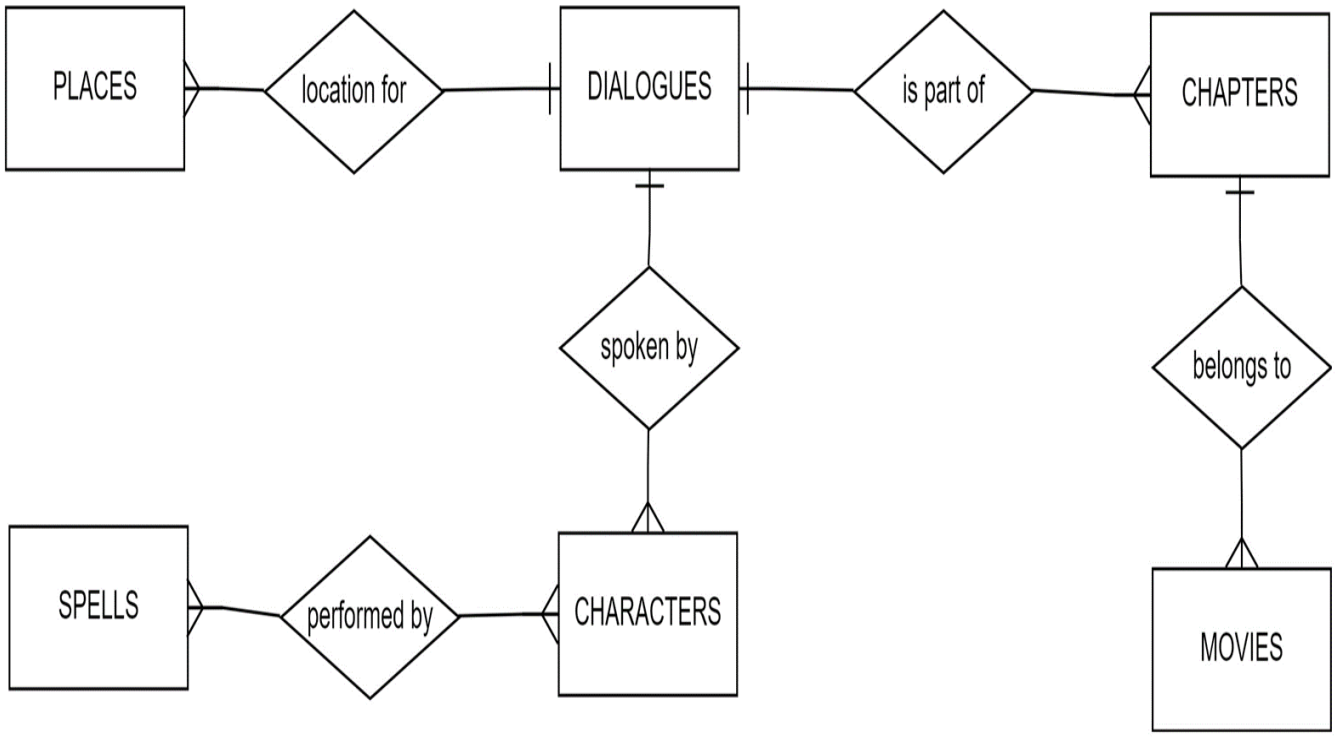
This is our report about the work we have done to build the final project for the Database and Data Mining course. It includes a description of the chosen dataset, details about our design decisions and an overview of the implementation steps for both the database itself and the queries that were written in SQL and Relational Algebra.

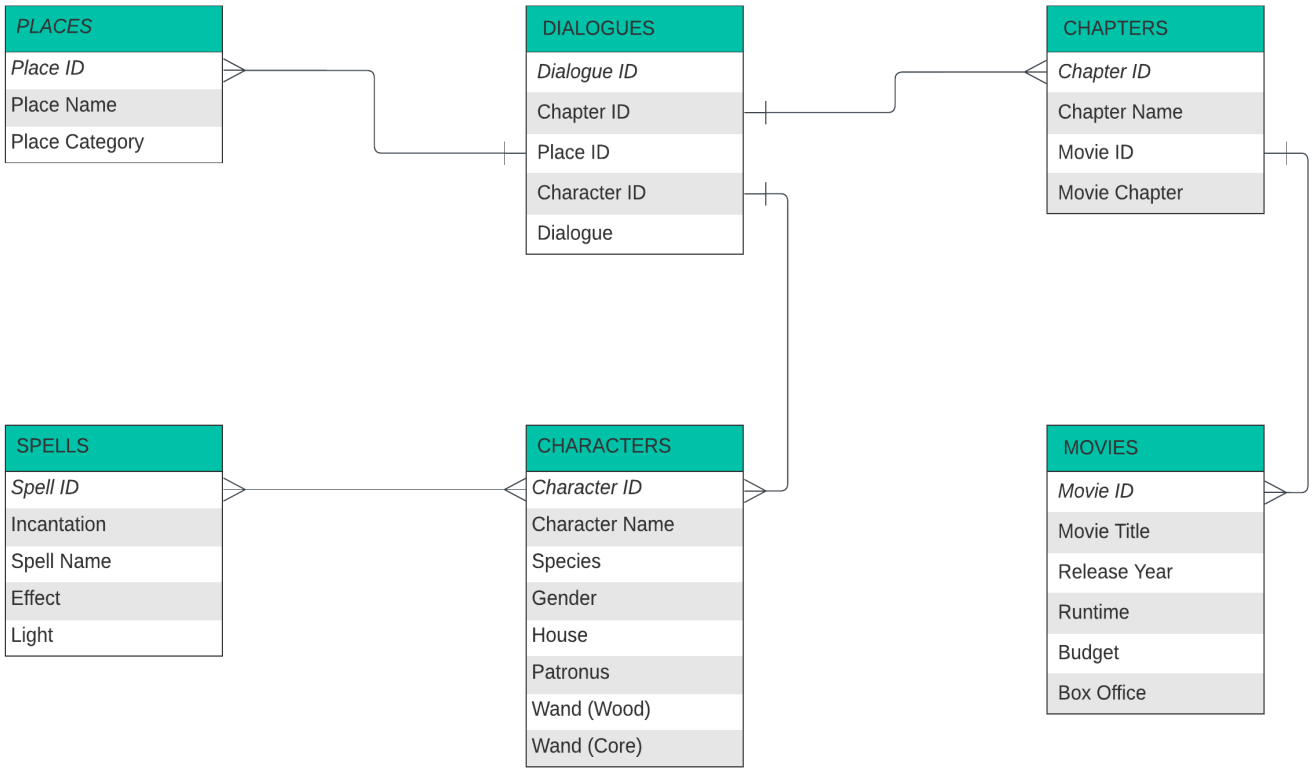
II. Information About The Dataset

The dataset we decided to work with contains data on the Harry Potter Movies. It can be obtained through Kaggle (https://www.kaggle.com/datasets/maricinnamon/harry-potter-movies-dataset), and it provides information about Movies, Characters, Dialogues, Chapters, Places and Spells.

III. Conceptual schema & ER Diagram

Unlike the ER Diagram, in the conceptual schema, we describe the relationships between entities using verbs.





For the design, we chose to mark the primary keys by writing them in *Italic*.

In this database schema for the Harry Potter movies, we have established several relationships between the tables:

The relationship between the **chapters** and **movies** tables is one-to-many. Many chapters belong to one movie, and this relationship is maintained through the Movie ID column in the chapters table, which serves as a foreign key referring to the Movie ID in the movies table.

The **dialogue** and **chapters** tables have a one-to-many relationship, where many dialogues are associated with one chapter. This association is facilitated by the Chapter ID column in the dialogue table, which is a foreign key referring to the Chapter ID in the chapters table.

A one-to-many relationship is established between the **dialogue** and **characters** tables. Many dialogues are spoken by a single character. This connection relies on the Character ID column in the dialogue table, which acts as a foreign key linked to the Character ID in the characters table.

The relationship between the **dialogue** and **places** tables reflects a one-to-many association, where many dialogues occur in a specific place. This link is maintained

through the Place ID column in the dialogue table, which references the Place ID in the places table. It assists in categorizing dialogues by their respective locations.

The relationship between the **spells** and **characters** tables represents a many-to-many connection. Many characters can perform many spells, and this complex relationship is typically managed through a junction table.

These relationships serve to organize and connect the various elements of the Harry Potter movies dataset, allowing for structured data storage and retrieval based on these associations.

IV. Implementation of The Database

With the dataset and the ER diagram built, we implemented the actual database.

First, we created all the entity tables with their respective attributes and primary keys (as designed in the ER diagram). Then, we imported the csv files and set the respective primary keys.

V. Populating The Database

After building all the tables with the command CREATE, we imported them as csv files. We have created a file called: table creation.sql.

VI. SQL Queries

These are the questions that we were curious to find an answer about in our database:

1. What is the average runtime of movies released before 2005?
2. Which are the top 5 characters with the most dialogues?
3. Which is the number of dialogues for each character in The Chamber of Secrets movie?
4. Which are the chapters where both Harry Potter and Hermione Granger have dialogues?
5. Which are the names of characters who have not spoken in chapters that take place in the "Corridors" location?

We translated the Queries in SQL and in Relational Algebra. The SQL implementation of the queries can be found in queries.sql.

VII. Conclusion

We built a database through which we can analyze data on Harry Potter Movies. Thanks to this practical experience, which involved using various tools to address specific questions, we understood better the concepts covered in the lectures. We find this kind of project really useful for our future tasks.