ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Distributed
Computing

HS 2020

Prof. R. Wattenhofer
Simon Tanner

Computational Thinking
Exercise 8 (Data and Storage)

1 Journal Article Database

The relation depicted in the following table contains information on articles published in scientific journals. Each article has a unique identifier (ID), a title, a name, issue and page number of the journal within which it was published. Also, if results of an article previously appeared in a technical report, the unique identifier (TR-ID) of this technical report is given. Each journal is published once a year. A new article always starts on a new page within a journal. You can assume that the endpage of an article is never smaller than the startpage. Answer the following questions based on the data in the table below.

ID	title	journal	issue	year	startpage	endpage	TR-ID
42	Cuckoo Hashing	JAlg	51	2004	121	133	87
33	Deterministic Dictionaries	JAlg	41	2001	69	85	62
33	Deterministic Dictionaries	JAlg	41	2001	69	85	56
39	Dictionaries in less space	SICOMP	31	2001	111	133	47
57	P vs NP resolved	JACM	51	2008	1	3	99
77	What Gödel missed	SICOMP	51	2008	1	5	98
78	What Gödel missed	Nature	2222	2008	22	22	98

a) A *superkey* is a set of columns whose entries uniquely identify every row in the table. A *candidate key* is a minimal superkey, i.e. if you remove any column from the candidate key, you will not be able to uniquely identify every row.

Which of the sets below are candidate keys for the given table?

```
1. {ID}
2. {ID,TR-ID}
3. {title}
4. {title,year,journal}
5. {ID,title,TR-ID}
6. {startpage,journal,issue}
```

b) Which of these SQL statements are invalid?

```
1. SELECT * FROM Articles WHERE endpage - startpage > 10;
2. SELECT * FROM Articles WHERE endpage - startpage < 0;
3. SELECT SUM(title) FROM Articles;
4. SELECT AVG(year) FROM Articles WHERE title LIKE 'C%';
5. SELECT COUNT(*) FROM Articles GROUP BY year;
6. SELECT year, COUNT(*) FROM Articles WHERE COUNT(*) > 10 GROUP BY year;
```

c) How many rows does this query return for the given table?

```
SELECT DISTINCT ID FROM Articles WHERE year < 2006;
```

2 Database Design

Swiss cat and dog owners are required by law to microchip their pets. This greatly improves the chances of succesful pet discovery in case the pet has run away or was stolen. The Animal Identity Service (ANIS) is a company that operates the Swiss animal database and allows veterinarians, animal shelters and police stations access if they are trying to find owners of lost pets. Unfortunately, the ANIS website was vulnerable to SQL injection attacks and an evil attacker managed to delete the entire database including all backups. Let's help them reconstruct their lost data!

- For each animal ANIS wishes to provide name and microchip id
- Every animal has exactly one owner. For each owner ANIS wishes to provide a name, an address, a phone number, and an email address
- To ensure the safety of a found animal at a shelter, ANIS provides additional information by recording any number of allergies the animal may have. We store a unique identifier for every possible allergy including its name and a number from 1 to 5 which indicates how strongly the allergy in question afflicts the animal.
- To help protect against animal microchip tampering and replacement, ANIS provides information about the type (cat, dog or other) and about any possible breed (Rottweiler, Norwegian Forest Cat, etc.) of the animal in question (note that some animals may belong to multiple breeds). For every distinct animal type or breed we store a unique identifier including the type or breed's name.

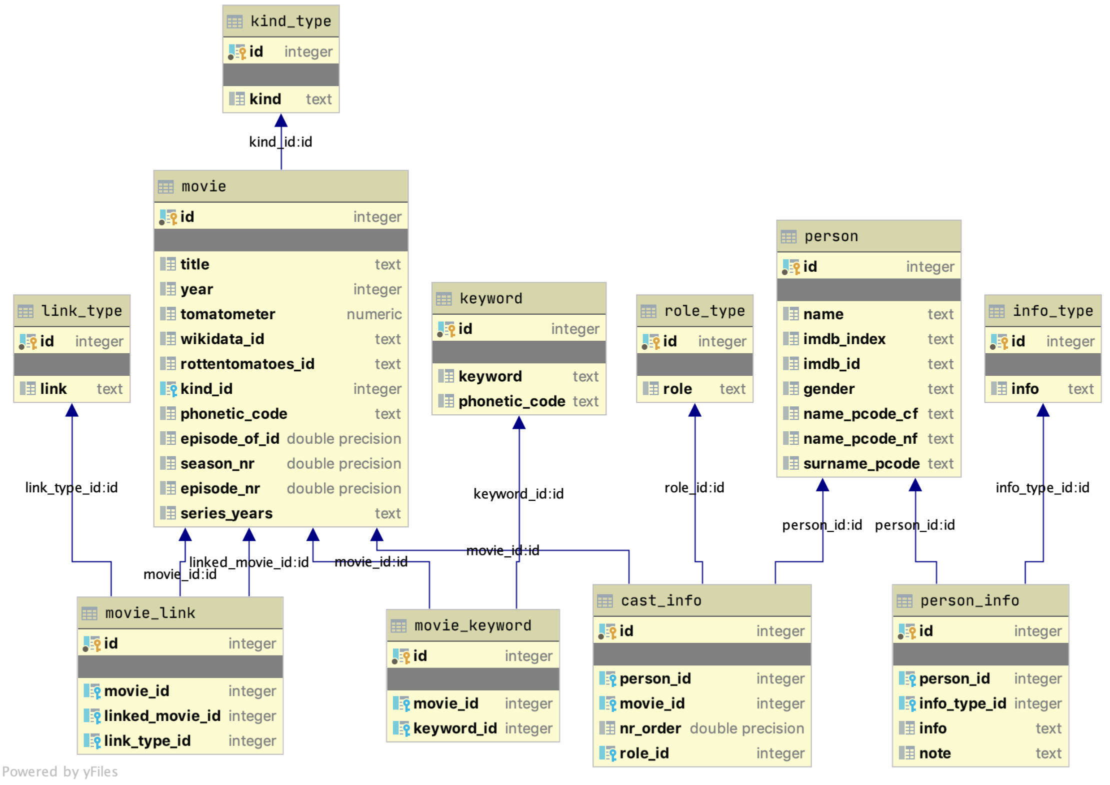
Draw an Entity Relationship Diagram for the data set described above. Make sure that you specify all cardinality constraints and that the resulting tables do not contain redundant entity sets, relationships, or attributes. Please include all made assumptions in your answer. Don't forget to assign primary key and foreign key constraints and mark attributes as unique wherever necessary

3 Database Queries

Take a look at the movie database schema to write SQL queries for the following questions. You can try out your queries in the notebook to this exercise or by connecting to the PostgreSQL database from a terminal with the psql client with user *student* and password *disco*2020:

→ notebook

```
psql -h disco-db.tik.ee.ethz.ch -U student -d moviedb_v3
```



a) Display the title and id of the first five entries in the movie table.

b) Display the first two rows of the movie table after sorting it by title in alphabetically reversed order.

c) How many movies came out since the year 2000?

d) What's the tomatometer score of the movie titled 'The Matrix'?

e) How many movies have a higher tomatometer score than 'The Matrix'?

f) Display the 5 years in which the average tomatometer score of movies was the best. Show this average score, too.

g) Which movie titles start with a capital X? List them in reverse alphabetical order.

h) How many movie titles contain the string 'fight'?

4 Advanced Database Queries

You will need to use table joins to answer the following queries for the movie database of the previous question. To find the first movie ’Steve Buscemi’ was cast in, you could use a query like:

```
SELECT movie.title, movie.year
FROM cast_info
JOIN person ON cast_info.person_id = person.id
JOIN movie ON movie.id = cast_info.movie_id
WHERE person.name = 'Buscemi, Steve'
ORDER BY year ASC
LIMIT 1;
```

a) List all actresses who were cast in ’The Matrix’.

b) How many female directors can you find in the database?

c) List all directors who have directed more than 20 projects and have also appeared at least once as an actor in any project (movie, tv show, etc).

d) Which movie is described by the most keywords?

e) What’s the average, maximum and minimum number of keywords per movie?

f) List all actors with an average tomatometer score over 85% who were cast in at least 30 movies and who appear in at least one movie made since 2000.

g) List all directors who have achieved to direct at least 10 movies with a tomatometer score of over 90%.