DIAS: Data-Intensive Applications and Systems Laboratory
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne
Building BC, Station 14
CH-1015 Lausanne
URL: http://dias.epfl.ch/

Databases Project – Spring 2017

Prof. Anastasia Ailamaki

Team No: 15

Names: Colin Branca, Jules Courtois, Yoan Martin

Contents

DIAS: Data-Intensive Applications and Systems Laboratory
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne
Building BC, Station 14
CH-1015 Lausanne
URL: http://dias.epfl.ch/

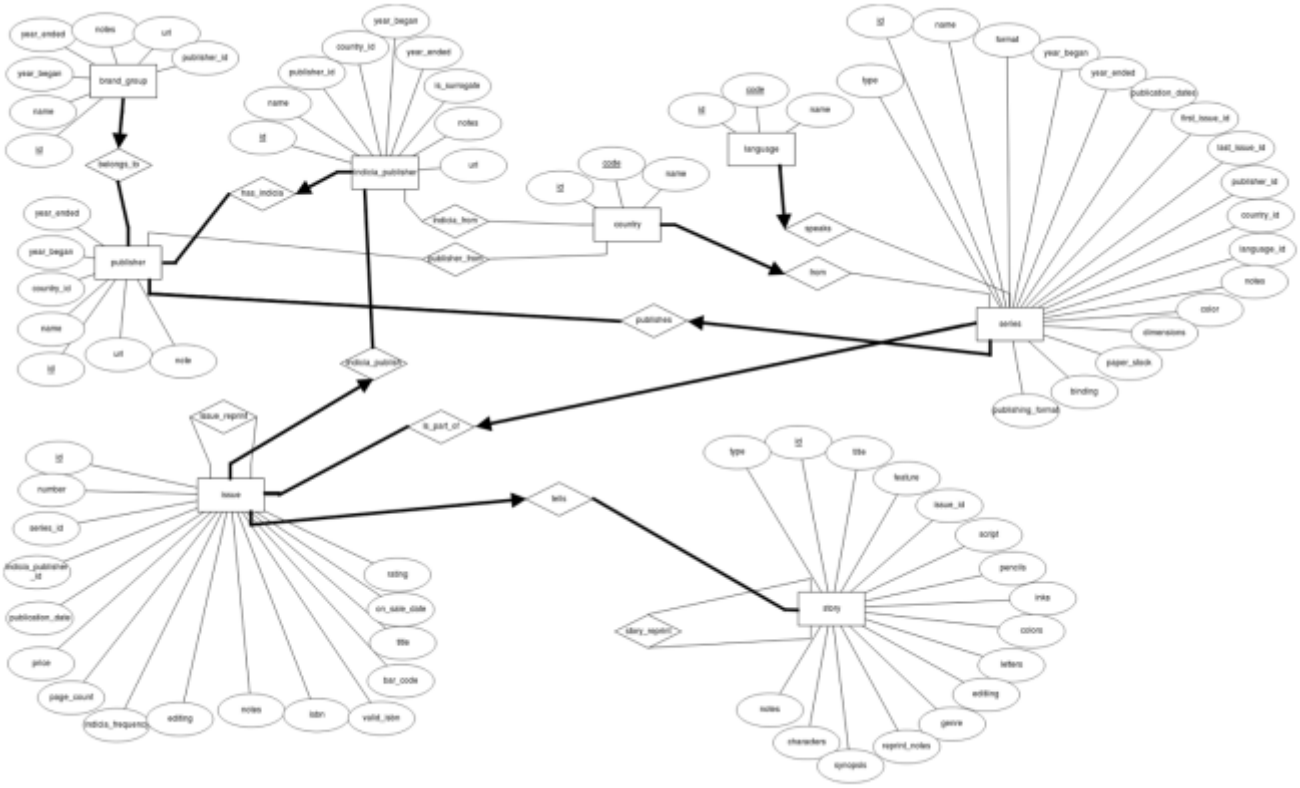
Deliverable 1

Assumptions

When we create the tables, we assume that each CHAR has a max size. No longer elements will be added later.

Entity Relationship Schema

Schema



Description

There are two main parts in the schema: Books with Story, Issue and Series and Company with Publisher, Indicia Publisher and Brand Group. Let's first describe each part before explaining the connections between them.
Books: This part is describing a book in the more general sense. The physical book is an issue. It tells a story and it is part of a series. For example, Harry Potter is a famous series. Harry Potter and the Philosopher's Stone is a story from this series. Finally, the book with ISBN X is a physical book telling this story. Since these three components are highly connected. There are a lot of constraints between them.

DIAS: Data-Intensive Applications and Systems Laboratory
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne
Building BC, Station 14
CH-1015 Lausanne
URL: http://dias.epfl.ch/

1) An issue tells at least a story.
2) A story must be told by exactly one issue.
3) Each series has at least one issue.
4) An issue is part of exactly one series.

Company: This part is describing the company which publish some books. The main company is the publisher. It has smallest companies, the indicia publishers and it holds some brand group. For example, Marvel is the general company. It holds a brand group called Thor and the physical books are published by an indicia publisher, Thor Entertaining Group, which is part of Marvel. As for the book part, the company part has also a lot of constraints.

1) Each publisher has at least one brand group.
2) Each brand group belongs to exactly one publisher.
3) Each publisher has at least one indicia publisher.
4) Each indicia publisher is part of exactly one publisher.

Then, there are some connections between these two parts. The publisher publishes a series. It means that the publisher creates a series but the issues are printed by the indicia publisher. Here are the constraints related to these connections.

1) A publisher publishes at least one series.
2) A series is published by exactly one publisher.
3) An indicia publisher publishes at least one issue.
4) An issue is published by exactly one indicia publisher.

Finally, there are some smallest relationships. Firstly, the publisher, the indicia publisher and the series come from a country and the series have a language. Secondly, there is a reprint relationship between two stories or two issues. We consider only two constraints.

1) A publisher, indicia publisher or series comes from exactly one country.
2) A series has exactly one language.

N.B. We do not create a story_type and series_publication_type entity since these data contain only one attribute if we exclude the id. To simplify the schema, we add an attribute type containing the name of the type in the Story/Series entity.

Relational Schema

ER schema to Relational schema

To create the table, we simply took each entity or each relationship and we translate them into a table. We took care to respect the constraints using NOT NULL or PRIMARY KEY to state the “at least” constraint and we permit each “at most” relation to appear in only one column in the connected table.

DIAS: Data-Intensive Applications and Systems Laboratory
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne
Building BC, Station 14
CH-1015 Lausanne
URL: http://dias.epfl.ch/

DDL

```sql
CREATE TABLE StoryReprint (
  id INT,
  origin_id INT NOT NULL,
  target_id INT NOT NULL,
  PRIMARY KEY (id)
);

CREATE TABLE IssueReprint (
  id INT,
  origin_issue_id INT NOT NULL,
  target_issue_id INT NOT NULL,
  PRIMARY KEY (id)
);

CREATE TABLE Country (
  id INT,
  code CHAR(4) NOT NULL,
  name CHAR(36) NOT NULL,
  PRIMARY KEY (id),
  UNIQUE (code),
  UNIQUE (name)
);
```

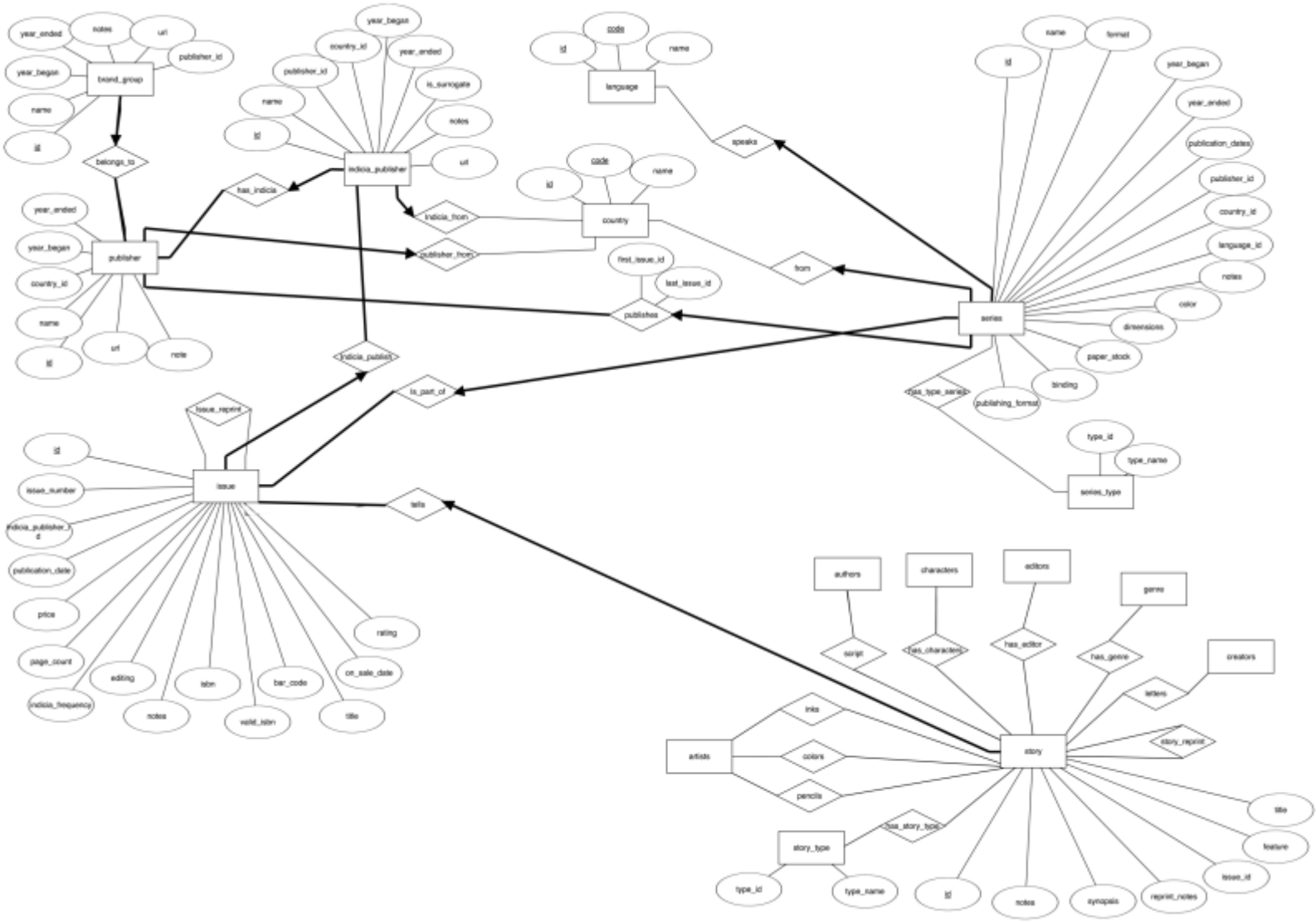
```sql
CREATE TABLE Story (
  id INT,
  type VARCHAR(50) NOT NULL,
  title VARCHAR(50) NOT NULL,
  feature VARCHAR(50),
  issue_id INT NOT NULL,
  script VARCHAR(50),
  pencils VARCHAR(50),
  inks VARCHAR(50),
  color VARCHAR(50),
  letters VARCHAR(50),
  editing VARCHAR(50),
  genre VARCHAR(50),
  characters VARCHAR(50),
  synopsis VARCHAR(150),
  reprint_notes VARCHAR(50),
  notes VARCHAR(50),
  PRIMARY KEY (id),
  FOREIGN KEY (issue_id) REFERENCES issue (id)
);
```

For this deliverable, we design the schema together. Then we split the work individually. Colin Branca and Jules Courtois wrote the SQL commands to create the tables and Yoan Martin wrote this pdf document.

DIAS: Data-Intensive Applications and Systems Laboratory
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne
Building BC, Station 14
CH-1015 Lausanne
URL: http://dias.epfl.ch/

Deliverable 2

Schema



DIAS: Data-Intensive Applications and Systems Laboratory
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne
Building BC, Station 14
CH-1015 Lausanne
URL: http://dias.epfl.ch/

Description and Changes :

First of all, we had made a few mistakes in the direction and nature of the constraints which we have corrected, for example with the “Country” entity. Each Indica_Publisher or Series originates from exactly one country, whereas a country does not require to produce even one of either.

The other big change is around the “Story” entity : we have created numerous new entities and relations which represent some of the attributes in the original database like the people who worked on each story based on their role, the genre of the story, or the characters featured in it. At first, we divided the people into artists, authors, editors and typesetters, for which we created the corresponding N-to-N relations as seen on the diagram. But after some reflexion and looking more into the data, we've realized that some authors are also artists or editors, and some artists are also typesetters which creates a duplication of the data. We are looking into changing this for the next version of the project, maybe by creating a “People” table and using ISA relationships.

To parse the original data and create the new tables, we had to write extra scripts in Python which automatically turned the story table into multiple smaller tables, and removed duplicates.

Data Loading

Query Implementation

```
/*a) Print the brand group names with the highest number of Belgian indicia publishers */
Select BG.name
From BrandGroup BG, (
  Select COUNT(distinct *)
  From Publisher P, (
    Select *
    From IndiciaPublisher IP, Country C
    Where IP.country_id = C.id and C.name = 'Belgium')
  Where P.id = IP.publisher_id)
Where BG.publisher_id = P.id and P.ROWNUM = 1
```

```
/*b) Print the ids and names of publishers of Danish book series*/
SELECT P.id, P.name
FROM Publisher P, Series S, Country C
WHERE S.country_id = C.id and C.name = 'Denmark' and S.publisher_id = P.id
```

DIAS: Data-Intensive Applications and Systems Laboratory
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne
Building BC, Station 14
CH-1015 Lausanne
URL: http://dias.epfl.ch/

```
/*c) Print the names of all Swiss series that have been published in magazines */
SELECT S.name
FROM Series S, Country C, Serie_type T, Has_Serie_Type H
WHERE H.serie_id = H.type_id and Serie_type.name = 'magazine' and S.country_id = C.id and C.name = 'Switzerland'

/*d) Starting from 1990, print the number of issues published each year*/
Select *
From Issue I
Group By I.publication_date
Where I.publication_date >= 1990

/*e) Print the number of series of each indicia publisher whose names resembles 'DC comics'*/
Select Count(distinct *)
From Issue I, Series S
Where (I.id = S.first_issue_id or I.id = S.last_issue_id) and I.Indiciapublisher_id IN(
  Select IP.id
  From IndiciaPublisher IP
  Where IP.name LIKE '%DC comics%')

/*f) Print the titles of the 10 most reprinted stories */
Select title
From (
  Select S.title AS title, COUNT(DISTINCT SR.origin_id) AS count_reprint
  From Story S, StoryReprint SR
  Where S.id = SR.origin_id
  Group By S.id, S.title
  Order By count_reprint DESC)
Where ROWNUM <= 10

/*g) Print the artists that have scripted, drawn, and colored at least one of the stories they where involved in */
Select A
From Artists A, Story S, Inks I, Colors C, Pencils P
Where A.id = I.artist_id and S.id = I.story_id and A.id = C.artist_id and S.id = C.story_id and A.id = C.artist_id and S.id =
C.story_id
```

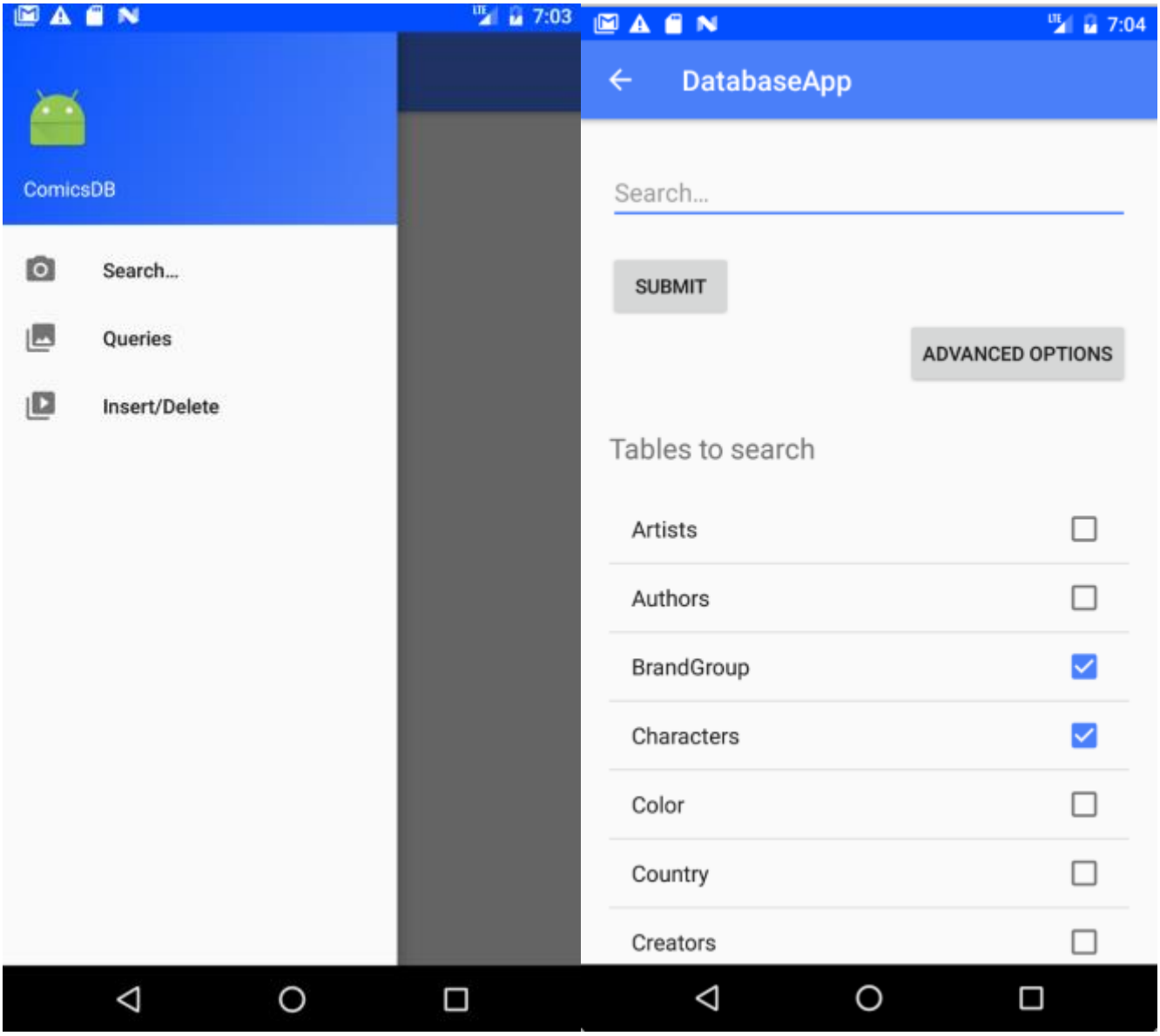
DIAS: Data-Intensive Applications and Systems Laboratory
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne
Building BC, Station 14
CH-1015 Lausanne
URL: http://dias.epfl.ch/

```
/*h) Print all non-reprinted stories involving Batman as a non-featured character */
Select S
From Characters C, Has_charaters HC (
  Select S
  From Story S, StoryReprint SR
  Where not exists (select * from  StoryReprint SR where SR.origin_id = S.id)
)
Where C.name != 'Batman' and C.id = HC.character_id and S.id = HC.story_id
```

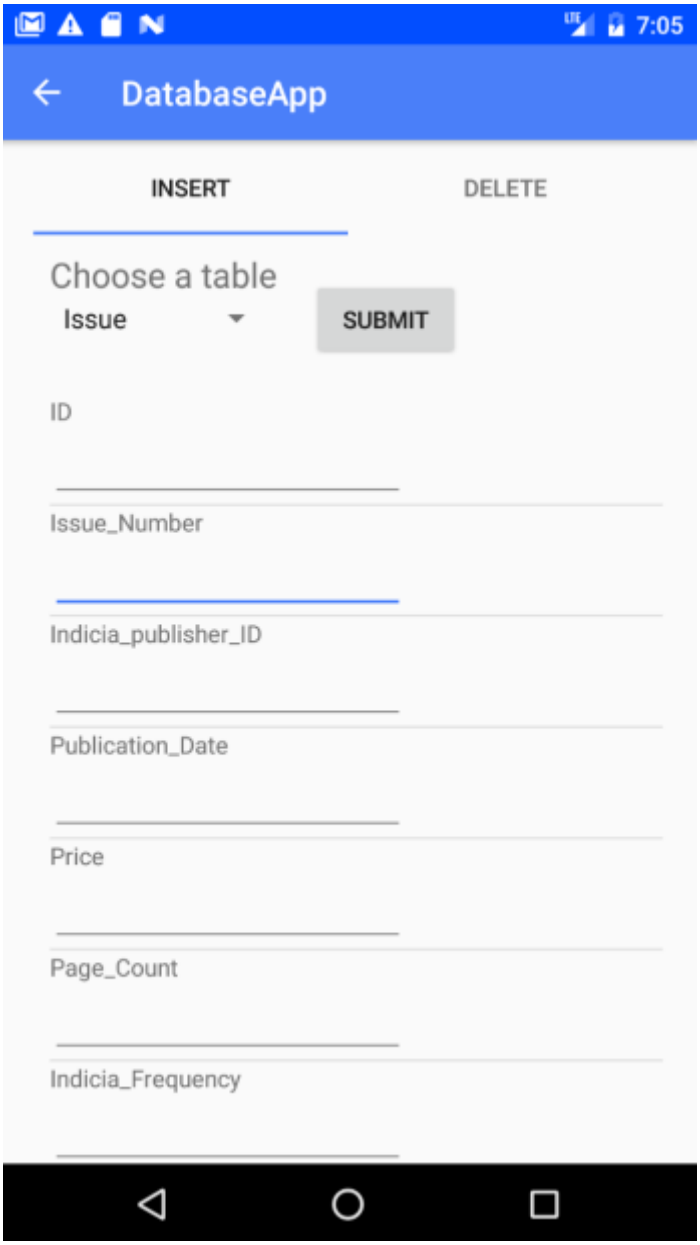
Interface

For the interface, we use an android application. The navigation in the app is entirely made using the top-left menu button. The user is able to select what he wants to do with the database and he is redirected to the corresponding activity. The frontend part of the application is working well. Unfortunately, we encounter problems with the database connection. So the biggest part of the work for the last deliverable is to work on the backend.

DIAS: Data-Intensive Applications and Systems Laboratory
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne
Building BC, Station 14
CH-1015 Lausanne
URL: http://dias.epfl.ch/



DIAS: Data-Intensive Applications and Systems Laboratory
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne
Building BC, Station 14
CH-1015 Lausanne
URL: http://dias.epfl.ch/



DIAS: Data-Intensive Applications and Systems Laboratory
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne
Building BC, Station 14
CH-1015 Lausanne
URL: http://dias.epfl.ch/

General Comments

We have met some difficulties due to the data format, for example in dates and how to compare them, or inconsistencies in the names of the authors or even the actual initial database. This will require some work before having a fully functional import.

Concerning the work attribution, Yoan Martin did the Interface, Jules Courtois coded some scripts to clean the dataset and Colin Branca corrected the work done for the last milestone and wrote the SQL queries.