CSCE100 Introduction to Informatics
Fall 2019

Programming Assignment 5: Hello Database

Points: 145 points. Assignment Date: October 15, 2019 Due Date: October 17, 2019

Objectives

1. To familiarize with MySQL queries and database
2. To prepare for the final class project through better understanding of the Chronicling America's repository—the database is a subset of the repository

Additional Resources

Official MySQL Tutorial: https://www.tutorialspoint.com/mysql/index.htm
SQL Tutorial: https://www.w3schools.com/sql/

Problem

A database has been pre-built for you. For this Assignment, you are required to access the database using MySQL queries to understand better the database. The database is a subset of the Chronicling America's repository, which is maintained by the Library of Congress. Each entry in the database is the metadata (attributes) of a newspaper page that has been archived and scanned into a digital form. The database is made up of multiple tables, with each table capturing parts of the data. The relationship diagram of the tables is shown in Figure 1 below. For this assignment, your tasks are to carry out the following activities to familiarize yourself with accessing a SQL-based database.

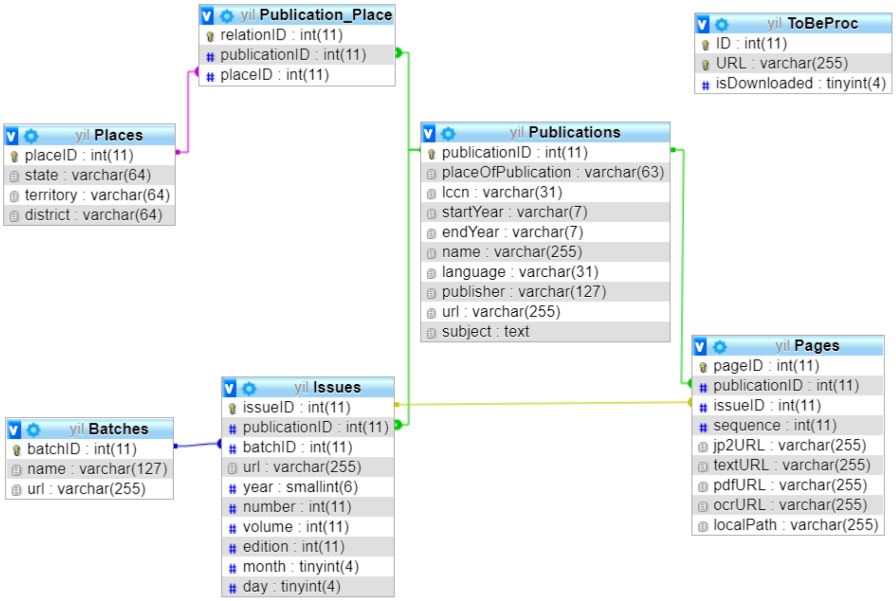

Figure 1. Entity-relationship (ER) diagram of the 7 tables used in the database. Take for example, the "Issues" table, it has a common link to the attribute "publicationID" with the "Publication_Place" table, and a common link to the attribute "batchID" with the "Batches" table.

STEP 1. Connecting Your CSE Account to MySQL

The Department of Computer Science and Engineering (CSE) provides a MySQL service for every user. To request a MySQL account on CSE:

1. Login to http://cse.unl.edu/account
2. On the left sidebar, click “Account Settings”
3. Check “MySQL Account?” and click "save settings"

Then, you will receive an e-mail on how to access MySQL.

To access the MySQL, there are two options:

1. **Through command line**: for example, in PuTTY, use the command: "mysql" to login to MySQL. *Example command*:

   ```
   mysql -u username -p database_name
   ```

   (*Note*: The password or the common alternative password character “*” will not show up on the prompt line.)
2. **Through phpMyAdmin**: a GUI MySQL database manager based on PHP, with the following URL: http://cse.unl.edu/phpMyAdmin/.
   Note: to execute any following MySQL query via phpMyAdmin, click “SQL” on the main menu.

STEP 2. Importing Database to Your Local CSE Account

To safeguard the database since it will be used by all students in this class, please download a copy of the database to your local CSE account where you will be able to access and modify it without interfering with other students' access.

1. Download assigned MySQL script, db_bak_8640_10-11-2019.zip, to your computer.
2. Log onto the MySQL database on your CSE account via http://cse.unl.edu/phpMyAdmin/
3. Click the database name as **same** as your **CSE account user name** on the left sidebar.
4. Click "Import" on the main menu.
5. Click "Choose File" button to choose the downloaded script downloaded in the first step on your local system.
6. Scroll down, and click "Go" button to import the script.

It will take a few moments for the download. Once the download is successfully completed, you will see a message on the screen that the "import has been successfully finished."

STEP 3. Using the “SELECT” Query to Retrieve Rows of Data from Database

The SELECT command is used to fetch data from the specified tables.

Syntax

```
SELECT column1, column2, ...
FROM table_name1, table_name2, ...;
```

Activity 3.1. To retrieve all data from the Publications table, type the following on the SQL workspace, and then click on the button ‘Go’:

```
SELECT *
FROM `Publications`;
```

Question 3.1. How many columns and rows are selected by the above query? (5 points)

Activity 3.2. To retrieve only the publication ID attribute (column) and the LCCN attribute (column) from the Publications table:

```
SELECT `publicationID`,`lccn`
FROM `Publications`;
```

Question 3.2. How many columns and rows are selected by the above query? (5 points)

Question 3.3. What is the difference between queries in Activities 3.1 and 3.2? (5 points)

STEP 4. Using the "WHERE" Query to Retrieve a Filtered Subset of Data Rows from Database

The WHERE clause is used to specify conditions in order to filter out the results from the SELECT command.

- A WHERE clause can be used also in DELETE and UPDATE command.
- Conditional signs include "Equal to" (=), "Not Equal to" (<>), "Greater than" (>), "Less than" (<), "Greater than or equal to" (>=), and "Less than or equal to" (<=).

Syntax

```
SELECT column1, column2, ...
FROM table_name
WHERE condition1 [AND [OR]] condition2...;
```

Activity 4.1. To retrieve all attributes (columns) of issues that are published later than 1900 from the Issues table:

```
SELECT *
FROM `Issues`
WHERE `year`>'1900';
```

Question 4.1. How many columns and rows are selected by the above query? (5 points)

Activity 4.2. To retrieve the issue ID attribute and the URL attribute of each issue that is published later than 1900 from the Issues table:

```
SELECT `issueID`,`url`
FROM `Issues`
WHERE `year`>'1900';
```

Question 4.2. What is the range of issue IDs in the queried result from Activity 4.2 above? (5 points)

Activity 4.3. To retrieve the issue ID attribute and the URL attribute of each issue that is published later than 1900 but earlier than 1930 from the Issues table:

```
SELECT `issueID`,`url`
FROM `Issues`
WHERE `year`>'1900'
AND `year`<'1930';
```

Question 4.3. How many issue IDs are retrieved by the above query in Activity 4.3? (5 points)

STEP 5. Using the "LIKE" Operator with WHERE to Filter Data Further

The (NOT) LIKE operator is used along with WHERE clause to filter out the results from the SELECT command. However, the difference between the (NOT) LIKE operator and "equal to sign" (=) or other comparative signs such as "bigger/smaller than" sign is that the LIKE operator allow a search of patterns. For example, a search of a value that contain character sequence "aha"; a search of a value that is start with "a"; a search of a value that is end with "z".

- You can use the (NOT) LIKE operator in place of the "equals to" sign.
- A condition with (NOT) LIKE operator can work with regular condition with the sign operators.
- You can use multiple conditions with (NOT) LIKE operator.
- In (NOT) LIKE operator, "%" represents zero, one, or multiple characters.
- In (NOT) LIKE operator, "_" represents a single character.

Syntax

```
SELECT column1, column2, ... FROM table_name
WHERE filed LIKE somevalue [AND [OR]] condition2;
```

Activity 5.1. To retrieve all rows of publications that started their publication in the 1900s from the Publications table (note the use of '%' in the query):

```
SELECT *
FROM `Publications`
WHERE `startYear` LIKE '19%';
```

Question 5.1. How many publications started their publication in the 1900s? (5 points)

Activity 5.2. To retrieve all rows of publications (and all their attributes) that did not start publication in the 1900s from the Publications table:

```
SELECT *
FROM `Publications`
WHERE `startYear` NOT LIKE '19%';
```

Question 5.2. How many publications are not started publishing in the 1900s? (5 points)

Activity 5.3. To retrieve all rows of publications (and their publication ID attribute and LCCN attribute) that started publication in the 1800s or the 1900s from the Publications table:

```
SELECT `publicationID`,`lccn`
FROM `Publications`
WHERE `startYear` LIKE '19%'
OR `startYear` LIKE '18%';
```

Question 5.3. How many publications started publication in the 1900s or the 1800s? (5 points)

Question 5.4. Why do we use the OR operator instead of the AND operator? (5 points)

Activity 5.5. To retrieve all rows of publications (and their publication ID attribute and LCCN attribute) that started publication in the 1960s or the 1990s from the Publications table (note the use of ‘_’):

```
SELECT `publicationID`,`lccn`
FROM `Publications`
WHERE `startYear` LIKE '196_'
OR `startYear` LIKE '199_';
```

Question 5.5. How many publications started publications in the 1960s or the 1990s? (5 points)

STEP 6. Using the "BETWEEN" Operator with WHERE to Filter Data Further

The (NOT) BETWEEN operator is used along with WHERE clause to filter out the results from the SELECT command. It filters values within a range.

- You can use the (NOT) BETWEEN operator in place of the "equals to" sign.
- A condition with (NOT) BETWEEN operator can work with regular condition with the sign operators.
- The BETWEEN operator is inclusive: begin and end values are included.
- The NOT BETWEEN operator is exclusive: begin and end values are not included.

Syntax

```
SELECT column_name(s)
FROM table_name
WHERE column_name BETWEEN value1 AND value2;
```

Activity 6.1. To retrieve all publications that started publication between 1901 and 1929 from the Publications table:

```
SELECT *
FROM `Publications`
WHERE `startYear` BETWEEN '1901' AND '1929';
```

Question 6.1. How many publications started publication between 1901 and 1929? (5 points)

Activity 6.2. To retrieve all publications that did *not* start publication between 1901 and 1929 from the Publications table:

```
SELECT *
FROM `Publications`
WHERE `startYear` NOT BETWEEN '1901' AND '1929';
```

Question 6.2. How many publications are not started publishing between 1901 and 1929? (5 points)

Activity 6.3. To retrieve all publications that did *not* start publication between 1901 and 1929, but *stopped* publication after 1970 from the Publications table:

```
SELECT *
FROM `Publications`
WHERE `startYear` NOT BETWEEN '1901' AND '1929'
AND `endYear`>'1970';
```

Question 6.3. How many publications started publication between 1901 and 1929, but stopped publication after 1970? (5 points)

Question 6.4. What are the names of the above publications? (5 points)

STEP 7. Using the "IN" Operator with WHERE to Filter Data Further

The (NOT) IN operator is used along with WHERE clause to filter out the results from the SELECT command. It filters the results with multiple specified values. It allows for **nested queries.** That is, the query result of a (inner) query can be used as the set of data for another (outer) query. It is a shorthand of multiple OR operations.

- You can use the (NOT) IN operator in place of the "equals to" sign.
- A condition with (NOT) BETWEEN operator can work with regular condition with the sign operators.
- A (NOT) IN operator allows nested query.

Syntax

```
SELECT column_name(s)
FROM table_name
WHERE column_name IN (value1, value2, ...);
```

OR

```
SELECT column_name(s)
FROM table_name
WHERE column_name IN (SELECT STATEMENT);
```

Activity 7.1. To retrieve all pages (and their attributes) that are published after 1890 from the Pages table:

```
SELECT *
FROM `Pages`
WHERE `issueID` IN (
  SELECT `issueID`
  FROM `Issues`
  WHERE `year`>'1890'
  );
```

Question 7.1. How many pages are retrieved using the above query? (5 points)

Question 7.2. Are all these pages published after 1890? Why? (5 points)

Activity 7.2. To retrieve all pages (and their attributes) that are *not* published after 1890 from the Pages table (not the 'NOT' operator):

```
SELECT *
FROM `Pages`
WHERE `issueID` NOT IN (
  SELECT `issueID`
  FROM `Issues`
  WHERE `year`>'1890'
  );
```

Question 7.2. How many pages are retrieved by the above query? (5 points)

Question 7.3. Are there more pages published before 1890 than after 1890? (5 points)

Question 7.4. Considering both queries in Activities 7.1 and 7.2, which query result include pages published in 1890? (5 points)

Activity 7.3. (IMPORTANT) To retrieve the name of all publications that are published in the state of California from the Publications table (note the nested-nested queries):

```
SELECT `name`
FROM `Publications`
WHERE `publicationID` IN (
  SELECT `publicationID`
  FROM `Publication_Place`
  WHERE `placeID` IN (
    SELECT `placeID`
    FROM `Places`
    WHERE `state`='California'
    OR `state`='california'
    OR `state`='CA'
    OR `state`='ca'
    )
  );
```

Question 7.5. How many publications are published in California? (5 points)

Question 7.6. Would the following query work? Why or why not? (5 points)

```
SELECT `name`
FROM `Publications`
WHERE `publicationID` IN (
  SELECT `publicationID`
  FROM `Places`
  WHERE `state`='California'
  OR `state`='california'
  OR `state`='CA'
  OR `state`='ca'
  );
```

Question 7.7. How to modify the query in Activity 7.3 to get the start year and end year of the publications that are published in California? (Write your query below) (10 points)

STEP 8. Using the "ORDER BY" Operator to Sort Data Returned by Query

The ORDER BY clause is used to sort the order of the results from the SELECT command.

- You can sort the results from the SELECT command on any **listed** field. You can sort the results on more than one field.
- You can use the keyword ASC or DESC to get results in ascending or descending order. By default, it is the ascending order.
- You can use WHERE clause in the regular way to add conditions.

Syntax

```
SELECT column1, column2, ...
FROM table_name
ORDER BY column1, column2, ... ASC|DESC;
```

Activity 8.1. To retrieve all publications that started publication in the 1900s from the Publications table, and order the query results by publication ID:

```
SELECT *
FROM `Publications`
 WHERE `startYear` LIKE '19%'
ORDER BY `publicationID`;
```

Question 8.1. What is the LCCN value of the smallest publication ID that started publication in the 1900s? (5 points)

Activity 8.2. To retrieve all publications (including all their attributes) that started publication in the 1900s from the Publications table, and order the results first by start year and then by publication ID:

```
SELECT *
FROM `Publications`
WHERE `startYear`
LIKE '19%'
ORDER BY `startYear`,`publicationID`;
```

Question 8.2. Look at the last row of the results retrieved by the above query. What is the place of publication? What is the LCCN value? When is the first year of publication? When is the last year of publication? What is the name of the publication? Which company published it? (10 points)

Question 8.3. Is the publication ID of the last row of the results the largest? Why or why not? (5 points)

Activity 8.3. To retrieve all publications (including all their attributes) that started publication in the 1900s from the Publications table, and sort the results by the LCCN attribute:

```
SELECT *
FROM `Publications`
WHERE `startYear` LIKE '19%'
ORDER BY `lccn`;
```

Question 8.4. How many publications are retrieved using the above query? (5 points)

Question 8.5. By default, MySQL uses the lexicographical order to sort strings. What is the correct order of the following strings (in ascending ordering): “a1”, “a10”, and “a2”? (5 points)

References

[1] "MySQL Tutorial", *www.tutorialspoint.com*, 2018. [Online]. Available: https://www.tutorialspoint.com/mysql/index.htm. [Accessed: 03- Aug- 2018].

[2] "SQL Tutorial", *W3schools.com*, 2018. [Online]. Available: https://www.w3schools.com/sql/default.asp. [Accessed: 03- Aug- 2018].

[3] University of Nebraska – Lincoln "Frequently Asked Questions - Unix-Linux", *Computer Science and Engineering*, 2018. [Online]. Available: https://cse.unl.edu/faq-section/unix-linux#node-302. [Accessed: 03- Aug- 2018].

[4] E. Lorang, L. Soh, Y. Liu, C. Pack and D. Rahimighazikalayeh, "Image Analysis for Archival Discovery (Aida)", *Projectaida.org*, 2018. [Online]. Available: http://projectaida.org/. [Accessed: 03- Aug- 2018].